Investigating viral parameter dependence on cell and viral life cycle assumptions

Carel Diederik Pretorius

School of Computational and Applied Mathematics

University of the Witwatersrand

A dissertation submitted for the degree of

Master of Science

March 2006

Declaration

Carel Diederik Pretorius,

I declare that this dissertation is my own, unaided work. It is being submitted for the Degree of Master of Science in the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.

Johannesburg, 2006

Abstract

This dissertation reviews population dynamic type models of viral infection and introduces some new models to describe strain competition and the infected cell lifecycle. Laboratory data from a recent clinical trial, tracking drug resistant virus in patients given a short course of monotherapy is comprehensively analysed, paying particular attention to reproducibility. A Bayesian framework is introduced, which facilitates the inference of model parameters from the clinical data. It appears that the rapid emergence of resistance is a challenge to popular unstructured models of viral infection, and this challenge is partly addressed. In particular, it appears that minimal ordinary differential equations, with their implicit exponential lifetime (constant hazard) distributions in all compartments, lack the short transient timescales observed clinically. Directions for future work, both in terms of obtaining more informative data, and developing more systematic approaches to model building, are identified.

Acknowledgements

I am indepted to many people and grateful for their contributions that have proven to be necessary for producing this dissertation. Firstly, I would like to thank my advisor, Dr. Alex Welte, for directing my interest in this field to a well-defined research project, and for helping me to develop the research skills necessary to produce this dissertation. I have relied on his knowledge of the field, insights for defining the scope of the project, and guidance for implementing various details. I enjoyed the opportunity of accompanying Dr. Welte on a visit to the University of California at Berkeley. This trip was made possible by a collaboration between SACEMA (South African Centre for Epidemiological Modelling and Analysis) and Dr. Wayne Getz at Berkeley. His hospitality made the trip memorable.

The laboratory of Lynn Morris at the National Institute for Communicable Diseases (NICD) has provided a Nevirapine resistance data set. Processing and analyzing of this data set has formed a crucial part of this project. I would like to thank Shayne Loubser, who has performed the laboratory work to produce the PCR data set, for many detailed discussions of PCR processes.

I would like to thank Roger Mateer for converting the Bayesian inference framework developed in this project, into an inference layer of DEBI, an in-house software product developed for simulations in biological and demographical problems. DEBI served as an independent check for some of my models, which were developed in Matlab. I also want to thank Robert De Mello Koch, Richard De Beer, Cobus van Eeden and Neil Pendock, who have made positive contributions towards this project, through discussions and exchange of ideas.

Finally, I would like to thank my family for their unwavering support and inspiration. I want to thank my father especially. Despite a serious health setback, he made time to be a willing proof reader of the manuscript, and a valuable advisor in general.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The purpose of this research project is to build mathematical models that can explain various important aspects of retro-viral strain competition, such as the emergence of resistant strains during antiviral treatment. The field of structured population dynamics [1, 2, 3], provides many methods and insights that can be applied to populations of interacting target cells, virions and infected cells. This project investigates ways of building structure into simple models governing dynamics at a population level. Models of biological systems generally result in a trade-off between mathematical tractability and biological faithfulness, as observed by [4]. Models that capture a great deal of biological detail can only be explored by computer simulation. Simple models allow mathematical analysis, but often oversimplify biological facts. Both approaches have limitations when applied to phenomena in biological systems. The principle adopted here is to make use of simple mathematical models, that nevertheless incorporate enough biological detail to explain clinically observed features.

Viral evolution is central to a number of clinically and epidemiologically relevant questions and is the subject of much ongoing research. A proper understanding of these phenomena is essential for optimal treatment planning, and can provide guidance for: 1) understanding the risk associated with non-adherence to a prescribed regimen, 2) designing regimen strategies, such interruption strategies of combination therapy, where different pharmacokinetics may lead to periods of monotherapy, 3) estimating the time period after which treating a general population with a particular regimen may become ineffective, and so on. It is studied here from the perspective of competition between viral strains in a host environment perturbed by anti-viral drug therapy, with particular emphasis on single dose Nevirapine treatment. Nevirapine is a non-nucleoside reverse transcriptase inhibitor and is often used in a resource-poor setting, to reduce the risk of transmission of HIV from mother to child. This treatment reduces the viral load of the mother whilst providing the infant with Nevirapine, via the placenta, at levels known to reduce the risk of viral transmission during labor. Researchers in this field have investigated strategies for reducing drug resistance in both mother and child. Data from a recent clinical trial, made available by the National Institute for Communicable Diseases (NICD), suggest that resistance to a single dose of Nevirapine treatment emerges within 14

days. This clinical outcome conflicts with predictions obtained from a well-known class of mathematical models [5, 6] widely used for treating in-vivo HIV dynamics. These models exhibit much slower response times. One of the aims of this study is to investigate appropriate modifications of these models.

Dynamic behavior of populations depends on how the individuals in that population are distributed with respect to the developmental variables of age, size, mass and so on. Most models for HIV dynamics assume exponential lifetime distributions for virions and target cells alike. An investigation of a particular model for in-vivo competition between strains, which assumes that cells and virions do not age or develop, shows that significant drug resistance emerges after three weeks of perfect treatment, even if the modelled fitness of the resistant strain is only marginally compromised. This is at odds with the clinical evidence which shows that resistance emerges on a time scale significantly shorter than three weeks. This disagreement would be even greater if the models incorporate the results of in-vitro experiments conducted to track the relative fitness of competing clones, which confirmed that there are significant fitness differences between competing strains [7].

Chapter 2 considers some of the widely used models for in-vivo HIV-1 dynamics i.e. those based on interacting populations of target cells and virions. Their assumptions of individual structure, for example age structure, are examined. They are found to be limited in addressing certain clinically relevant features. This is a motivation for investigating models with additional structure. It is a simple matter to produce a significant drug resistant virion population arbitrarily quickly by giving the lifetime distribution of productively infected cells an ever shorter tail. Gamma lifetime distributions are particularly easy to implement, by simply introducing a linear chain of compartments or stages [8, 9, 10]. Moreover, it has the exponential and delta distribution as two limiting special cases: An exponential distribution is a gamma distribution with one stage. A delta distribution, obtained in the limit as the number of compartments increase, models the assumption that all individuals live to the same age and then 'die'. Knowledge of the physiological mechanisms at the binding sites on the surfaces of cells and virions suggests constraints on how these models should depend on development or age. A simple algorithm suggests that cell and virion lifetimes follow a delta-like distribution, which means that the probability of survival is high for a characteristic period where after it rapidly expires. A system of partial and integro-differential equations is modified and implemented by means of a numerical scheme called the Escalator Boxcar Train (EBT) algorithm [11]. It is used to explore arbitrary age structure for populations of infected cells.

The resistance data consist of estimates of plasma viral load in blood samples collected over the space of one year, after the initial single dose of Nevirapine. Initial estimates for the concentration of K103N resistance variants have been reported by [12]. These estimates are produced by a Polymerase Chain Reaction (PCR) process, a standard method for amplifying and quantifying small quantities of DNA sequences. A rudimentary investigation of the PCR process is reported in **Chapter 3**. PCR is based on multiple cycles of enzymatic DNA synthesis that form a chain

reaction. Each cycle includes denaturation (splitting) of double stranded templates into complementary single strands, annealing of primers to their complementary sites on these single stranded templates, and extension of the primer. The efficiency of the process of synthesizing DNA is a critical factor in the attempt to estimate initial concentrations. Most methods are based on the assumptions that all templates are faithfully duplicated. Substantial effort was spent on developing methods to investigate the implications of PCR efficiency for accurate prediction of initial concentrations. The software was then used to estimate the relative fraction of two competing viral sub-populations. It was found that target specific amplification, as employed in the PCR process which produced the afore mentioned data set, is of limited usefulness in quantifying sub-populations of mRNA. An alternative, and possibly superior method is discussed as part of the chapter summary.

A quantitative theory of any physical system, such as host-viral dynamical systems, requires a description of all the uncertainties in the theory. These uncertainties are expressed as probability distributions over a corresponding parameter space. The model and parameter space have uncertainties associated with incorrect theories and inaccurate measurements respectively. The Bayesian approach is often used as a subjective interpretation of probability and is widely used to solve inverse problems. It is a systematic way of combining all the uncertainties and limitations in the forward modelling and measurement of a system [13]. A review of inverse problem theory is reported in **Chapter 4**, providing some theoretical background for the inference techniques used in this project. Two important discussions concern homogeneous and hierarchical prior distributions. The latter is often used when parameters in multiparameter inference problems are related in some way. For example, in a clinical trial designed to infer in-vivo viral parameters of different trial attendees, it is reasonable to assume that there exists a relationship between viral parameters of different individuals. Hierarchical Bayesian modelling is frequently applied to longitudinal data sets and clinical data [14, 15]. It has recently been used in the context of HIV dynamics [16, 17] and is applied in this project to infer parameters relating to the host-virus models developed in Chapter 2.

Bayesian inference of viral parameters is the subject of **Chapter 5**. When unrealistic assumptions about the lifetime distributions of infected cells are built into dynamical models, it leads to unrealistic inferred values for model parameters. The 'shape' of lifetime distributions of infected cells have been structured into a staging scheme. Inferring 'the number of stages' will shed some light on the tail of these distributions. The inference result can be obtained in a way that is insensitive to the precise value of the 'the number of stages'. This is achieved by means of a homogeneous prior distribution for the shape parameter, which assigns equal probability to a parameter corresponding to long-tailed (exponential), short-tailed (delta) and an intermediate (gamma-like) lifetime distribution. The resistance profiles, obtained from the PCR data set in Chapter 3, are too few to provide an accurate estimate of the population level distribution of viral parameters. The data set was not collected to meet our inference and modelling requirements. A positive outcome for work of

this kind would be a collaboration between clinicians and modelers, and a realization that mathematical modelling can provide valuable tools for the interpretation of clinical data.

Chapter 2

Population dynamical models for HIV infection

Dynamical processes that arise in epidemiology, ecology, pharmacokinetics, chemical engineering, and so on, can often be modelled as a flow of material between compartments, governed by mass balance laws. Compartmental modelling is a method for finding the differential equations that govern the flow or transfer of material between compartments.

The term 'material' could mean molecules, population members, target cells or virions, depending on the system being studied. 'Compartment' refers to a kinetically homogenous amount of material, in the sense that the material is assumed to be well mixed at all times and furthermore that it mixes instantaneously with new material entering the compartment [10]. All units experience the same probability per unit time of being transferred in a kinetically homogenous system. Furthermore, the transfer of material between compartments is modelled as a continuous process, determined by the transfer rates of the system. The transfer of material between compartments is in reality a discrete process, where an arbitrary small number of units is transferred in a (sufficiently) small time interval.

Bodily compartments for HIV infection would typically include lymphatic tissue, peripheral blood and so on. These compartments host the interaction between billions of target cells and virions and can be thought of as the limit of a corresponding stochastic system. This limit is obtained by enlarging the compartments until they contain enough individuals to smooth out fluctuations due to individual stochasticity. The compartmental models we build become *deterministic* in this limit and they govern the time evolution of the expectation values of the underlying *stochastic* processes. There are scenarios when this assumption may not hold, for example when viral loads are low in the early stages of infection or when suppressed during anti-viral treatment. Stochastic models for HIV infection are not investigated in this project.

Compartmental HIV models are set up as ordinary differential equations describing the time evolution of cell and virion cohorts. Populations of cells and virions are usually not differentiated beyond the viral genotype and mayor cell phenotype level. Individuals in such populations are assumed to be equal in every relevant respect,

for example: 1) they experience the same risk of dying or being cleared from the system, 2) they have same production rate, and 3) they have the same response to environmental factors, such as treatment. Models that incorporate various types of delays and physiological structure have been explored, providing a more nuanced description of viral load decline during therapy. Delays, for example, are used to model aspects of the life cycle of infected cells, such as the delay between viral cell entry and the onset of the production of new virions [18]. Partial differential equations [19, 20, 21] have been introduced to implement 'age structure' in models that include non-constant hazards of infection, production, survival and death. However, additional structure, whilst complicating both the models and their implementation, has yet to produce new insights for the management of HIV infection. From a modelling perspective, this chapter considers:

- What aspects of structured populations may explain the clinically relevant features of the system?
- How can these be modelled by extending simple models, for example, by introducing appropriate discrete categories of cell and virion populations?
- Which physiological insights can be promoted by mathematical modelling.

We introduce the following population structure, beyond those already used in the simplest models:

- Viral genotypes found: 1) circulating in bodily compartments, and 2) transcribed into the DNA of a host cell.
- Age of infected cells, where 'age' refers to time since infection by a viral genome.

The discussion proceeds by starting from the simplest models, which lack either of these population structures, and then considers each in turn, as well as models with both. The modelling principle of explaining known properties of the system by introducing only minimal extra degrees of freedom is applied.

2.1 A simple model for HIV infection

Perhaps the simplest sensible model of HIV infection is constructed by specifying time derivatives for the populations of healthy T cells $T(t)$, 'productively infected' T cells $P(t)$ and free virions $V(t)$. The following simple dynamical model is widely used in this field and has been thoroughly reviewed [5, 6]:

$$\frac{dT(t)}{dt} = S_T - kV(t)T(t) - \mu_T T(t) \tag{2.1}$$

$$\frac{dP(t)}{dt} = fkV(t)T(t) - \mu_P P(t) \tag{2.2}$$

$$\frac{dV(t)}{dt} = N\mu_P P(t) - \mu_V V(t) \tag{2.3}$$

where,

- The body produces new T cells at a rate S_T. Production will typically be assumed to be constant over the course of a simulation. In reality there is no doubt an effect from chronic infection, which is not properly understood.
- μ_T and μ_P are constant mortality rates for healthy and productively infected T cells respectively.
- Healthy cells are homogenously mixed with virions and they are infected at rate k. This mass action term for the infection rate is probably not physically precise, but may be adequate as an average of the mixing in crucial compartments on the time scale being probed.
- A fraction of f cells entered by a virion become productively infected.
- Infected cells produce N new virions over their lifetime. A constant fraction of new virions are able to reproduce, i.e they are equipped with functional reproduction, capsid and other proteins.
- Virions have a constant hazard of 'death' μ_V, and this term can be increased to include the loss of virions that manage to enter and infect a cell.
- This simple mathematical model mimics a biological description of the viral life cycle. More details are given in Appendix A.

This class of models implements the assumption that: 1) u_P can be described by an exponential distribution with mean $\frac{1}{u_P}$, i.e. that infected cells do not 'age', and 2) there is no delay between the infection and viral production stages in the life cycle of an infected cell.

Equilibrium state

There exists a trivial equilibrium condition for the model of equations (2.1)-(2.3), when the system is in a non-infected state:

$$T_{eq} = \frac{S_T}{\mu_T} \tag{2.4}$$

$$P_{eq} = 0 \tag{2.5}$$

$$V_{eq} = 0 \tag{2.6}$$

It is easy to find the equilibrium condition during infection by setting the time derivatives of all state variables equal to zero [6]:

$$T_{eq} = \frac{\mu_P \mu_V}{fk\beta} = \frac{T_0}{R_0} \tag{2.7}$$

$$P_{eq} = (R_0 - 1)\frac{\mu_T \mu_V}{fk\beta} \tag{2.8}$$

$$V_{eq} = (R_0 - 1)\frac{u_T}{fk} \tag{2.9}$$

where,

- $\beta = N\mu_P$ represents the average number of virions produced by a productively infected cell.
- $T_0 = \frac{S_T}{\mu_T}$ is the initial uninfected T-cell population.
- $R_0 = \frac{S_T \beta k}{\mu_T \mu_P \mu_V}$ represents the basic reproductive value of the virus [6]. Note that a disease steady state can only exist if $R_0 > 1$. This can also be stated as $k > \frac{\mu_T \mu_P \mu_V}{S_T \beta}$, highlighting k, the infection rate.
- The equilibrium state is stable [22], i.e the system tends (to return) to its equilibrium value when perturbed from it.

We now provide a short account of some of the developments of HIV dynamics modelling and we focus on those that have clear origins in clinically observed features of HIV infection. More detailed accounts can be found in [5, 6]. The most important impact on our understanding of HIV dynamics to be derived from a mathematical model, resulted from the seminal work of Ho et al [23]. Fitting a linearized version of the model given by equations (2.1)-(2.3) to the viral load data for 20 patients during anti-viral treatment, elucidated the rapid turnover and relatively short average lifetime of populations of infected cells. This high turnover rate is part of the mechanism by which the virus extends its genetic diversity. This finding, together with his pioneering work in HIV treatment, has won David Ho Time's award: "Person Of The Year" in 1996. HIV dynamics modelling has not continued to produce such dramatic findings. Nevertheless, attempts to capture clinically observed features of HIV infection has led to important modelling questions, and the remainder of this section highlights some of these.

An observed feature during anti-viral treatment is that patients experience a rapid decline in viral load, but not quite viral eradication. When treatment is modelled as a reduction in infectivity, this observed fact can only be produced for a narrow range of modelled parameters. This is called 'fine tuning'; model parameters have to be tuned very carefully for the model to produce a certain relevant behavior. 'Fine tuning' problems often point to an incomplete model; a physical mechanism is missing from the model, and it would not be necessary to 'fine tune' the parameters of a complete model. This is one of the motivations for including additional structure in simple models for HIV dynamics.

Clinical data show that plasma viral loads for infected individuals follow a rapid rise that peaks within a few weeks after infection. Viral loads then decline, reaching a set point which is very low compared to the peak and which varies substantially between patients. Characteristics of primary infection, including: 1) the time to peak 2) amplitude of the peak, and 3) eventual viral set point, have been correlated with subsequent disease progression. The main conclusion is that a high viral set point after primary infection can be correlated with increased risk during disease progression.

The characteristic decline has been attributed to: 1) target cell limitation, and 2) immune system responses. The extent to which these processes contribute to the decline in viral load, has not been answered completely. Staffard et al [24] have shown that variation in the parameters of model (2.1)-(2.3) can account for the variation in the rise and subsequent fall of viral loads observed between patients. They also show that the viral load in many patients falls to levels below what can be predicted by this simple model. Including immune system responses into model (2.1)-(2.3), does allow the prediction of low viral set points, but [24] concludes that it is not clear how the contributions of different processes in viral load decline may be disentangled.

Including additional physiological processes

It is a simple matter to incorporate additional physiological processes, by adding new population compartments and dynamical terms to the model of equations (2.1)-(2.3), depending on the features to be addressed. Some of the familiar extensions are briefly discussed below:

- T cells undergoing clonal expansion to maintain homeostasis of total T cell populations, or in response to antigen threats, can be captured by adding a logistic term to (2.1):

$$\frac{dT(t)}{dt} = S_T + T(t)\left(1 - \frac{T(t)}{T_{max}}\right) - kV(t)T(t) - \mu_T T(t) \qquad (2.10)$$

- Some models use functional forms such $S_T(a, V, t)$ to model the fact that both aging and viral infection tend to reduce thymus activity [25]. The thymus, an organ situated behind the breast plate, expresses most 'self' (as opposed to foreign) epitopes. The thymus helps to select self-tolerant T cells and to de-select T cells expressing high affinity for 'self' epitopes. The self-tolerant T cells are initially 'naive', and continue to circulate through various bodily compartments. Maturing is reached after these 'naive' T cells are exposed to their cognitive epitopes, leading to 'memory' T cells. Lymphocyte circulation models are often captured by distinguishing compartments of 'naive' and 'memory' T cells, with thymus activity typically modelled as input for 'naive'

T cell compartments [25]. For the purposes of this investigation we assume that $S_T(t) = S_T$ is constant.

- Patients do not experience viral eradication during $HAART$ (highly active antiviral treatment), even when the ability of the virus to infect is almost completely suppressed. Instead, clinical evidence suggests that viral load declines in more than one phase. This feature can be explained by including other viral compartments [6], that cannot be reached by treatment, of cells being 'latently' infected and living for much longer than the more active 'productively' infected cells. These cells may be reactivated to produce new viral particles, feeding new virions into this peripheral blood system. This is an example of how a delay process can be captured by introducing a suitable compartment.

- A minimal model for HIV in vivo dynamics can easily be extended to keep track of more than one viral strain [6], and such a model is the subject of the next section.

It seems reasonable to expect that additional population compartments and dynamical terms will extend the range of dynamical behavior that can be modelled. What is not clear is how additional dynamical behavior can be explained by introducing minimal complexity, or degrees of freedom, into a system such as the one given by equations (2.1)-(2.3).

2.2 Multi strain models

A minimal model of HIV infection by N_s strains is:

$$\frac{dT(t)}{dt} = S_T - T(t)\sum_{i=1}^{N_s} k_i V_i(t) - \mu_T T(t) \tag{2.11}$$

$$\frac{dP_i(t)}{dt} = f k_i V_i(t) T_i(t) + T(t)\sum_{j\neq i}^{N_s} \varepsilon_{ji} k_j V_j(t) - \mu_{P_i} P_i(t) \tag{2.12}$$

$$\frac{dV_i(t)}{dt} = N_i \mu_{P_i} P_i(t) - \mu_{V_i} V_i(t) \tag{2.13}$$

where,

- P_i is a categorization of T cells infected with genome of viral strain V_i.

- Each infected cell P_i will produce N_i virions during its life time.

- f is the probability of error free transcription and ε_{ji} is the probability of a transcription error of the required kind, to cause a cell entered by a virion carrying a genome of strain j to become infected with a genome of strain i. This could be as high as about 0.0001 for strains differing by a single base mutation, and rapidly declines as the strains become more distinct.

From a modelling point of view it is not necessary to include more than two viral strains when studying strain competition during monotherapy. The wild type strain will refer to the fittest drug sensitive strain, whilst the drug resistant strain will refer to the fittest drug resistant strain. If the drug sensitive wild-type strain is eradicated during anti-viral therapy, then any less fit drug sensitive strain will also be eradicated. For the remainder of this chapter, (V_1, P_1) and (V_2, P_2) will refer to virions and cells productively infected with the drug sensitive wild type and the drug resistant strain respectively. V_1 and V_2 will refer $K103$ and $K103N$ when the modelling ideas of this chapter is connected with clinical data in later chapters. $K103N$ is the fittest Nevirapine resistant mutant [26].

Equilibrium State

The equilibrium state of a two strain model is obtained by means of perturbation analysis (Appendix B). A perturbation analysis is set up to make small corrections to the equilibrium state of a model without mutation, a state which is easy to obtain. Equilibrium states are given by:

$$T_{eq} = \frac{\mu_P}{fk_1} \tag{2.14}$$

$$P_{1_{eq}} = \left(\frac{S_T - \mu_T T(0)}{k_1 T(0)}\right) - \epsilon\left(\frac{k_2}{k_1} P_2^{(1)}\right) + O(\epsilon^2) \tag{2.15}$$

$$P_{2_{eq}} = \epsilon\left(\frac{k_1 P_1(0) T(0)}{\mu_P - f k_2 T_0}\right) + O(\epsilon^2) \tag{2.16}$$

where ϵ is the mutation rate, which is assumed to be the same for both strains.

Some important features of multi strain models

Important features of multi strain models include:

- Immune system pressure, a key determinant in 'fitness' measures, is not included explicitly in this model. It is modelled indirectly by means of clearance rates μ_X, $X = T, P$ and V. Equilibrium conditions are also dependent on immune system pressure, and in particular on the degree to which the immune response is strain-specific or cross-reactive [27]. Section 2.5 shows how a immune system response can be modelled by a interplay between maternity (virion production) and mortality of infected cells.

- There is no equilibrium between strains of varying fitness that do not mutate into each other; without mutation the fittest strain will outcompete all less fit strains, and a marginal relative weakness leads to extinction.

  'Fitness' is meant to be a measure of the reproductive ability of a strain. Data on relative *in vivo* fitness is scarce, and are generally based on *in vitro* growth competition experiments between mutant clones. In [7] fitness is equated with ability to reproduce in a defined host cell environment.

- The time required to approach equilibrium between identically fit strains which differ by a single mutation is many years.
- Resistance to monotherapy, often conferred by single base mutations, requires three weeks of almost perfect selective pressure to emerge at levels comparable to pre-treatment levels of the wild-type strain.

The last point is at odds with clinical evidence, as indicated by analysis of the data made available by $NICD$. This analysis suggests a more rapid emergence of resistance during single dose Neverapine treatment, even when these models are modified to include significant fitness differences for viral strains (as suggested by *in vitro* tests). Age-structured models are now introduced, which helps us to interpret resistance data in the light of assumptions about physiological structure in the population of infected cells.

2.3 Age structured models for HIV infection

Introducing age structure is a natural way of generalizing HIV models and allows for variable hazards of infection, recovery, reproduction, death and other individually variable attributes. The focus here is to demonstrate that the turnover rates of competing populations, such as strains of varying drug sensitivity, depend not just on the mean, but also the shapes of lifetime distributions in the model.

The following model, similar to the one first introduced by [19] and more recently by [20], describes time dependent age distributions for productively infected T cells $P(a,t)$.

$$\frac{dT}{dt} = S_T - kV(t)T(t) - \mu_T T(t) \tag{2.17}$$

$$\frac{\partial P(a,t)}{\partial t} = -\frac{\partial P(a,t)}{\partial a} - \mu_P(a)\, P(a,t) \tag{2.18}$$

$$\frac{dV}{dt} = \int_0^\infty m(a)\, P(a,t)\, da - \mu_V(a)\, V(a,t) \tag{2.19}$$

$$P(0,t) = fkV(t)T(t) \tag{2.20}$$

where,

- Construction of a specific initial value problem involves specifying $T(t_0)$, $P(a,t_0)$, and $V(t_0)$, where $T(t_0)$ and $P(a,t_0)$ are initial concentration of healthy and infected T-cells prior to infection, and $V(t_0)$ the initial virus concentration.
- Boundary conditions are given by equation (2.20). Once appropriate initial conditions have been specified, equations (2.17) to (2.20) can be used to determine $T(t)$, $P(a,t)$, and $V(t)$ for $t > t_0$.
- The introduction of f captures the possibility that only a fraction of cells removed from the healthy cell pool, would become productively infected.

- Infected cells $P(a,t)$ experience age dependent 'mortalities' $\mu_P(a)$.
- $m(a)$ is the rate of virion production for an infected cell aged a. Estimates for $m(a)$ are the subject of [20]. Two hypothetical functional forms for $m(a)$ are studied in Section 2.5.
- This model would be substantially complicated by the introduction of an age structure for the population of healthy T cells. It is a simple matter to include an age structure for V, the population of virions, in order to model infectivity k as a function of virion age. However, virions have life spans that are very short compared to other time scales in this system and an age structure for virions may typically be omitted.

Equilibrium condition

Obtaining the equilibrium condition is a simple matter that 1) begins with setting the time derivatives in the model of equations (2.17) to (2.19) to zero:

$$T_{eq} = \frac{s_T}{u_T + kV_{eq}} \tag{2.21}$$

$$P_{eq}(a) = P_{eq}(0)l(a) = fkV_{eq}T_{eq}l(a) \tag{2.22}$$

$$V_{eq} = \frac{1}{\mu_V}\int_0^\infty m(a)P_{eq}(a)\,da \tag{2.23}$$

and 2) makes use of the fact that when a population $p(a,t)$ experiences time independent but age dependent mortality rates $\mu(a)$, then $p(a,t)$ obeys:

$$p(a,t) = p(0,t-a)e^{-\int_0^a \mu(s)\,ds} \tag{2.24}$$

$$= p(0,t-a)l(a) \tag{2.25}$$

where $l(a) = e^{-\int_0^a \mu(s)\,ds}$ is the probability that an individual will survive from birth to age a. The equilibrium condition for the model of equations (2.17) to (2.20) follows, after applying (2.22) to (2.23), as the solution of a simple set of simultaneous equations [20]:

$$T_{eq} = \frac{\mu_V}{fkN} \tag{2.26}$$

$$P_{eq}(a) = \left(s_T f - \frac{\mu_T\mu_V}{kN}\right) l(a) \tag{2.27}$$

$$V_{eq} = \frac{s_T f N}{\mu_V} - \frac{\mu_T}{k} \tag{2.28}$$

where $N = \int_0^\infty m(a)l(a)\,da$ is the 'burst size' of an infected cell. This equilibrium condition is stable, as is the case with the equilibrium condition of a comparable system of ordinary differential equations [20].

An infected steady state will exist only if $V_{eq} > 0$, i.e. if:

$$\frac{fkN}{\mu_V} > \frac{\mu_T}{s_T} = \frac{1}{T_0} \tag{2.29}$$

The ratio of infected cells before infection to infected cells at equilibrium is given by:

$$\frac{T_0}{T_{eq}} = R_0 \tag{2.30}$$

where R_0, the 'basic reproductive number' of an infected cell, equals the number of infections of healthy cells due to one infected cell that is introduced to the initial phase of the infection. The simplicity of this expression is the result of a naive model for viral dynamics, which does not, for example, include clonal expansions of T cells in response to various threats. Substituting (2.26) into (2.30), shows that the expression

$$R_0 = \frac{T_0 fkN}{\mu_V} \tag{2.31}$$

summarizes the fitness of the virus [21]. It is interesting to investigate a more general scenario, where the survivorship of infected cells is dependent on viral production schedules:

$$l(a) \quad = \quad e^{-\int_0^a \mu_P(m(a))\, da} \tag{2.32}$$

In this case infected cell mortality $\mu_P(m(a))$ is an implicit function of age. Following [21, 28], Section 2.5 investigates how this implicit relationship can be constrained by physiological arguments relating to the life-cycle of infected cells.

2.4 Multi-strain age structured models for HIV infection

The age structured model given by (2.17)-(2.20) can be generalized to model strain competition:

$$\frac{dT(t)}{dt} = S_T(t) - T(t)\sum_{i=1}^{N_s} k_i V_i(t) - \mu_T T(t) \tag{2.33}$$

$$\frac{\partial P_i(a,t)}{\partial t} = -\frac{\partial P_i(a,t)}{\partial a} - \mu_{P_i}(a) P_i \tag{2.34}$$

$$P_i(0,t) = f k_i V_i(t) T_i(t) + T(t)\sum_{j\neq i}^{N_s} \varepsilon_{ji} k_j V_j(t) \tag{2.35}$$

$$\frac{dV_i(t)}{dt} = \int_0^\infty m_i(a) P_i(a,t)\, da - \mu_{V_i} V_i(t) \tag{2.36}$$

The equilibrium state can again be derived by means of perturbation analysis involving the mutation rate ϵ (Appendix C):

$$
\begin{aligned}
T_{eq} &= \frac{\mu_V}{fkN_1} && (2.37)\\
V_{1eq} &= \left(\frac{S_T f N_1}{\mu_V} - \frac{\mu_T}{k}\right) - \epsilon\left(\frac{k_2}{k_1}V_2^{(1)}\right) + O(\epsilon^2) && (2.38)\\
V_{2eq} &= \epsilon\left(\frac{k_1 N_2 V_1(0) T(0)}{\mu_V - f k_2 N_2 T(0)}\right) && (2.39)\\
P_{1eq}(a) &= \left(f k_1 V_{1eq} T_{eq} + \epsilon k_2 V_{2eq} T_{eq}\right) l_1(0,a) + O(\epsilon^2) && (2.40)\\
P_{2eq}(a) &= \left(f k_2 V_{2eq} T_{eq} + \epsilon k_1 V_{1eq}(t) T_{eq}\right) l_2(0,a) + O(\epsilon^2) && (2.41)
\end{aligned}
$$

Simplified lifetime distributions

Various assumptions can reasonably be made about the age dependence of virion production and cell mortality rates. For example, models with constant mortality assume that cells do not age, and hence that the lifetime distribution follows an exponential distribution. When the lifetime distribution is given by the delta function, i.e, when it is sharply peaked around the mean life expectancy for infected cells, the assumption is made that all cells live a hazard free life, up to a certain age, and then 'die'. It could also be imagined that the probability of cell death is given by a gamma distribution, which is often used to represent a system with multistage internal processes. The lifetime of such a system can be seen as the time delay between the first and last sub-processes and 'death' occurs only after a number of sub-processes are completed.

These assumptions can be built into simple delay models, which can then be used to model simplified age distributions. Consider a modification of the original constant hazard model. By making the lifetime distribution in the P population perfectly sharply peaked about its mean value λ, we obtain a fixed delay model:

$$
\begin{aligned}
\frac{dT(t)}{dt} &= S_T(t) - T(t)\sum_{i=1}^{N_s} k_i V_i(t) - \mu_T T(t) && (2.42)\\
\frac{dP_i(t)}{dt} &= f k_i V_i(t) T_i(t) + T(t)\sum_{j\neq i}^{N_s} \varepsilon_{ji} k_j V_j(t) - f k_i V_i(t-\lambda) T_i(t-\lambda) && (2.43)\\
&\quad -T(t-\lambda)\sum_{j\neq i}^{N_s} \varepsilon_{ji} k_j V_j(t-\lambda)\\
\frac{dV_i(t)}{dt} &= \frac{\beta_i P_i(t)}{\lambda} - \mu_{V_i} V_i(t) && (2.44)
\end{aligned}
$$

where,

- β_i is the rate of viral production for infected cell of type P_i over its fixed lifetime of λ.

Exponential and delta-like lifetime distributions lead to very different dynamical behavior. By means of Figure 2.3, it will be demonstrated shortly how resistance emerges more rapidly when the lifetimes follow a delta distribution. A more realistic alternative to discrete delays is a continuous distribution of delays, and finding simple ways of modelling such delays by introducing suitable compartments into the model. Delay models arise in dynamical systems when the current state of the system is dependent on the history of the system [29]. A model with delays can always be expanded into a model without explicit delay terms, by including the process causing the delays [10]. Gamma lifetime distributions are particularly easy to implement, by simply introducing a linear chain of compartments or stages. Moreover, it has the exponential and delta distribution as two limiting special cases: An exponential distribution is a gamma distribution with one stage. A delta distribution captures the assumption that all individuals live to the same age and then 'die'. It is obtained in the limit as the number of compartments increase. The 'linear chain trick' has been applied in [8], which explores some of the implications that the assumed lifetime distribution of infected cells may have for inferences based on viral load data.

Thus, implementing a gamma lifetime distribution amounts to introducing a number of categories, or life stages:

$$\frac{dT(t)}{dt} = S_T(t) - T(t)\sum_{i=1}^{N_s} k_i V_i(t) - \mu_T T(t) \tag{2.45}$$

$$\frac{dP_{i,1}(t)}{dt} = f k_i V_i(t) T_i(t) + T(t)\sum_{j\neq i}^{N_s} \varepsilon_{ij} k_i V_i(t) - \mu_{Pi,1} P_{i,1}(t) \tag{2.46}$$

$$\frac{dP_{i,2}(t)}{dt} = \mu_{Pi,1} P_{i,1}(t) - \mu_{Pi,2} P_{i,2}(t) \tag{2.47}$$

$$\vdots$$

$$\frac{dP_{i,n}(t)}{dt} = \mu_{Pi,n-1} P_{i,n-1}(t) - \mu_{Pi,n} P_{i,n}(t) \tag{2.48}$$

$$\frac{dV_i(t)}{dt} = N_{i,1} P_{i,1}(t) + N_{i,2} P_{i,2}(t) + \ldots + N_{i,n} P_{i,n}(t) - \mu_{V_i} V_i(t) \tag{2.49}$$

where,

- Each infected cell population P_i, is stepped through n stages: $P_{i,1}$ to $P_{i,n}$.
- The mean (of the exponentially distributed) time spent in each stage is $\frac{1}{\mu_{Pi,j}}$.
- Mortality rates $\mu_{Pi,n}$ are adjusted in each stage to give overall mortality μ_P.
- The stage in the infected cell life cycle when viral production starts, can be controlled by setting $N_{i,j} = 0$ for some of the stages $j <= n$.

The subsystem governing P_i can be decoupled from T and V_i by cutting of the supply of new infected cells. This is achieved by setting by setting $k_i = 0$ in (2.46). The infected cells P_i are then governed by the following system of linear ODE's:

$$\begin{bmatrix} P'_{i,1}(t) \\ P'_{i,2}(t) \\ \vdots \\ P'_{i,n-1}(t) \\ P'_{i,n}(t) \end{bmatrix} = \begin{bmatrix} -\mu_1 & 0 & \dots & \dots & 0 \\ \mu_{i,1} & -\mu_{i,2} & \dots & \dots & 0 \\ \vdots & & \ddots & & \vdots \\ \vdots & \dots & \mu_{i,n-2} & -\mu_{i,n-1} & \vdots \\ 0 & \dots & \dots & \mu_{i,n-1} & -\mu_{i,n-1} \end{bmatrix} \begin{bmatrix} P_{i,1}(t) \\ P_{i,2}(t) \\ \vdots \\ P_{i,n-1}(t) \\ P_{i,n}(t) \end{bmatrix}$$

stated compactly as:

$$P'_i(t) = \mathbf{U}P_i \tag{2.50}$$

This system is easily solved for $P_i(t)$:

$$P_i(t) = e^{\mathbf{U}t}P_i(0) \tag{2.51}$$

where $P_i(0) = [P_{i,1}(0), P_{i,2}(0), \dots, P_{i,n}(0)]$ is the initial population state. We consider a special initial condition $P_i(0) = [P_{i,1}(0), 0, \dots, 0]$ to observe survivorship in the cohorts $P_{i,n}$. Figure 2.1 demonstrates how mortality approaches a step function as the number of stages increase. The same (arbitrary) value for $P_{i,1}$ is used in each model.

Tab. 2.1: State variables and parameters

State Variable	Descriptions	
T	Uninfected, heathy T cell population	
$i = 1, 2$	Number of strains distinguished	
P_i	T cells infected with genome of strain V_i	
V_i	Infectious viral strain i	
V_1	Wild type	
V_2	Resistant type	
Parameter	**Descriptions**	**Value**
T_0	Initial T cell population	$9 \times 10^{10}\text{mm}^{-3}$
S_T	T cell production rate	$1,8 \times 10^{8}\text{mm}^3 d^{-1}$
μ_T	Mortality rate of healthy T cells	$\frac{1}{500} d^{-1}$
μ_{P_i}	Mortality rate for infected T cells	$\frac{1}{2.6} d^{-1}$
μ_{V_i}	Mortality rate for free virus	$\frac{1}{0.3} d^{-1}$
N_1	Number of free virions produced by P_1	500
N_2	Number of free virions produced by P_2	400
k_1	Rate at which T cells become infected by V_1	$4 \times 10^{-12}\text{mm}^3 d^{-1}$
k_2	Rate at which T cells become infected by V_2	$4 \times 10^{-12}\text{mm}^3 d^{-1}$
ϵ_{12}	Mutation rate from V_1 to V_2	$3 \times 10^{-5}\text{mm}^3 d^{-1}$
ϵ_{21}	Mutation rate from V_2 to V_1	$3 \times 10^{-5}\text{mm}^3 d^{-1}$

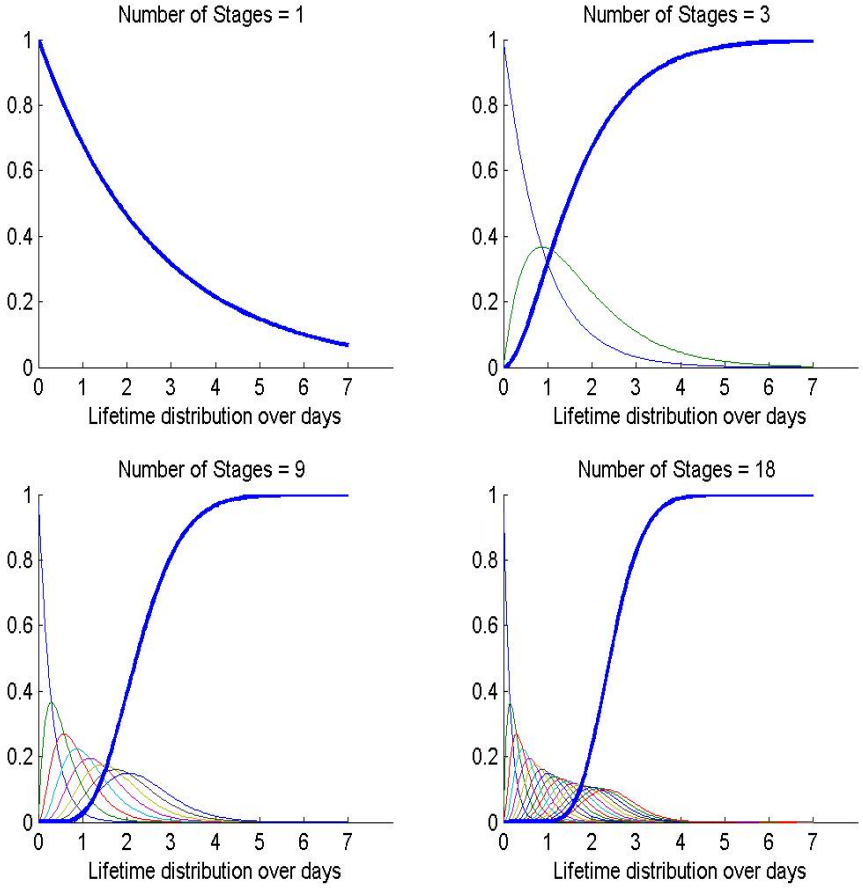

Fig. 2.1: Mortality or death can be equated with the rate at which the population exits the last stage, indicated by the increasing blue line.

The corresponding survivorship functions (in blue) and lifetime distributions (in red) are displayed in Figure 2.2. It can be seen that lifetime distribution becomes more peaked as the number of stages are increased. The survival function, which can be calculated from the rate at which cells exit the last stage, can be seen to approach a step function, as the number of stages are increased.

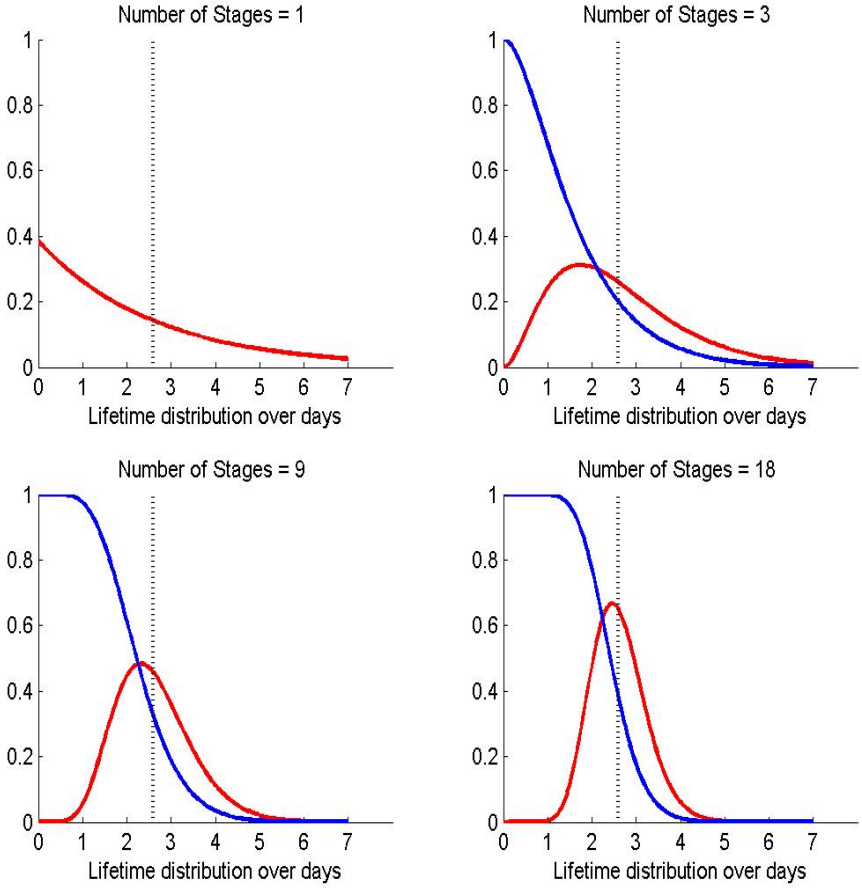

Fig. 2.2: Lifetime distribution (represented by the red curve) become sharply peaked as the 'number of stages' increase.

Figure 2.3 simulates 7 days of perfect treatment, by four different models. The same dynamical parameters, listed in Table 2.1, are used in each model. The only difference between these models is that the cohort of infected cells is stepped through an increasing number of stages.

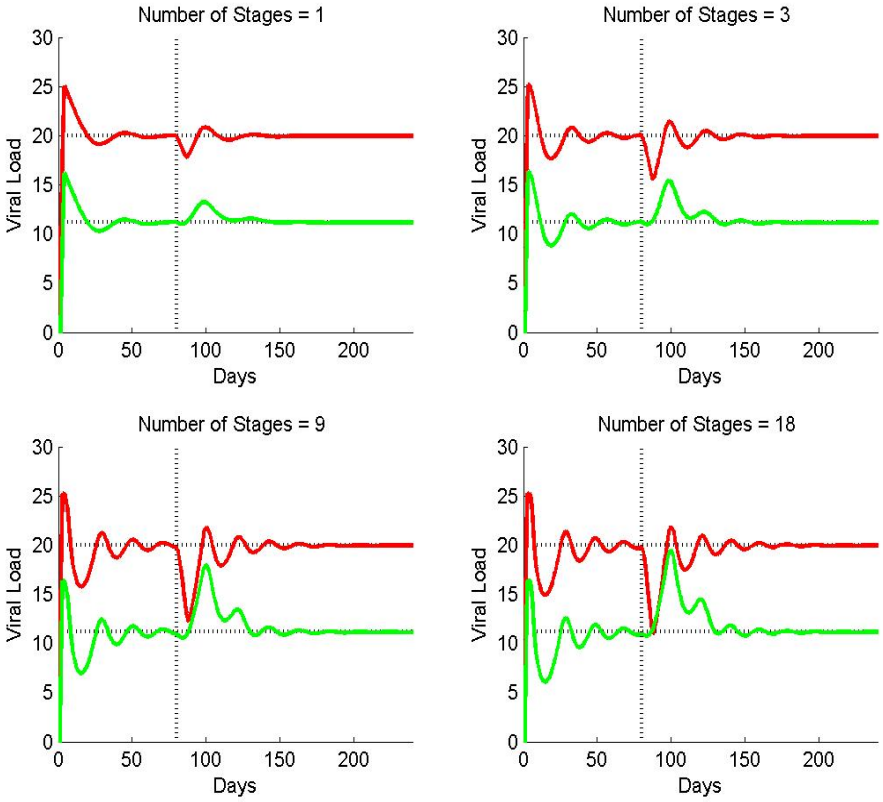

Fig. 2.3: Simulating 7 days of perfect RTI treatment of HIV, 80 days (indicated by vertical lines) after initial infection with one perfectly resistant(green) and one perfectly drug sensitive strain(red). All dynamical parameters are identical, except that the cohort of infected cells is stepped through an increasing number of stages. The key point is the differential impact on population turnover behavior. The horizontal line represents equilibrium states calculated with perturbation scheme (B.26) - (B.28), and it can be seen that the calculation is in good agreement with the simulated equilibrium state. Note that viral loads are displayed on a logarithmic scale.

We see that a resistant virion population emerges rapidly, when we assume a short tail for the lifetime distribution for infected cells. Thus, by varying the number of stages, and the exponentially distributed time spent in each stage, we can investigate a continuum of lifetime distributions. The overall distribution of the above mentioned scheme is a gamma distribution, and the exponential and delta distributions are two special cases. These two distributions will provide a 'natural' constraint for a range of possible distributions, and in the next section we look at ways to constrain lifetime distributions using physiological considerations.

2.5 Physiological constraints on 'age' dependent parameters for populations of infected cells

We have added age structure, and considerable extra complexity, into a minimally sensible model for multi strain HIV dynamics. This extra complexity is useful to explain certain clinically relevant features of host virus dynamics, such as the rapid emergence of resistance during treatment. It was shown that certain types of age structure can be modelled in a simple way by subdividing cell and virion populations through the use of discrete compartments or categories. The question arises whether the use of these simplifications can be justified on a physiological basis. This section attempts to show how physiological arguments can be used to suggest the shapes of the age distributions of populations of infected cells. These insights can then be used to inform or devise sensible discretization schemes.

Age dependent characteristics of an individual do not depend on age in an intrinsic way, but rather on internal development [9]. The 'age' of an individual happens to be highly correlated with development, and so are other attributes, including mass, size, and so on. In the context of virus and immune system dynamics, attributes include the number of functional glycoprotein 'spikes' on a virus particle, and the number of warning signals displayed by infected cells to immune system surveillance. Little is known about the details of how the viability of infected cells degrades as a result of their infected status and prevailing immune system responses. Four of the contributing processes are [6, 21, 30]:

- When a virus infects a cell it manipulates the biosynthetic machinery of the cell to produce viral proteins. The rate at which vital resources are lost is proportional to the rate at which viral proteins are produced. The extent to which the cell is compromised is called viral *cytopathicity*.

- These cells are fortunately equipped with a mechanism to inform the immune system that they have been hijacked to produce virions. The immune system reacts to eliminate the cell with a process called CTL-mediated lysis. The following is a most condensed account of this mechanism. A complete and accessible account is available in [31]. Some of the viral proteins are cut into peptides and transported, via the endoplasmic reticulum to the surface of the cell. Peptides of a certain length (approximately 9 amino acids), are loaded into the class I-MHC molecules on the surface, to inform killer T cells of its internal status.

- All cells experience a risk of being cleared by the system. This risk is independent of the cell's infection status and is called the 'background mortality'.

- Like all blood cells, T cells originate in the bone marrow where they descend from stem cells. Each T cell is at the end of a (possibly long) line of cell divisions, and the dependence of cell life expectancy on its position in this line is not explored here.

We now propose a model of infected cell mortality. To the extent that this model is realistic, infected cell lifetime distributions are quite sharply peaked and have rapidly decreasing tails. A number of physiological assumptions are built into the following model for a immune system detection mechanism:

- Viral protein fragments are loaded into the class I-MHC molecules at a rate which is proportional to viral production rate. These locations are uncorrelated and spread evenly over the cell membrane.
- There is a constant rate of class I-MHC inspection by cytotoxic T cells.
- Immune system surveillance 'samples' N_s class I-MCH molecules during each time unit.
- At least N_r of the 'sampled' class I-MCH molecules must display viral peptides for cross-linking to produce a 'kill this cell' type signal.
- The probability of survival of the cell is proportional to the probability that less than the required number of peptides will be available for cross-linking.

The process generating a 'kill this cell' signal is much more complicated than this algorithm may suggest. In reality, the distribution of these receptors over the cell membrane, situated as they are on mobile and floating structures, may only be approximately uniform. Accessory proteins associate with the heavy protein chains of these receptors, and serve to amplify the signal of antigen matching into the cytoplasm of the cell. The clustering of receptor and accessory proteins is called cross-linking, and it is vital for efficient antigen recognition. Cross-linking attracts nearby receptors to the linking site. This increases the concentration of receptors and amplifies the signal. Depending on the signal strength it generates, a receptor cluster thus formed interacts with other cell processes. Such processes are not included in this algorithm. Having pointed out some of the limitations of the algorithm, the remainder of this section aims to illustrate some of its implementations.

A number of physiological details must be specified before this model can be implemented:

- The original number of class I-MHC molecules is approximately 100000. The exact number is not directly relevant in this model, except that it provides an upper bound on sensible values of N_s. It is estimated that approximately 100 MHC molecules must be recognized by the receptors of a naive T cell to be activated [31]. The exact number is again not directly relevant, and serves as an upper bound for activation requirements. T cells that have already encountered their cognitive antigen, have less strict re-activation requirements.
- Two plausible viral production schedules are proposed by [20]. Both are characterized by a maximum production rate due to limited cellular resources.

They also take into account possible delays between initial infection and the onset of viral production.

The first functional form is a delayed exponential:

$$P(a) = \begin{cases} P_{max}(1 - e^{-\beta(a-a_1)}) & : \quad a \geq a_1 \\ 0 & : \quad \text{otherwise} \end{cases} \tag{2.52}$$

where a_1 is the earliest age of viral production and β controls the rate at which the maximum production rate (P_{max}) is approached.

The second is a Hill type function:

$$P(a) \quad = \quad P_{max}\left(\frac{a^n}{K_a^n + a^n}\right) \tag{2.53}$$

where K_a^n is a constant related to the half-suturation level and n an arbitrary constant.

- There is also no clear relationship between viral production rate and details of how fragments of viral proteins are loaded into class I-MHC molecules. The simplest assumption is that fixed number peptides (N_{pep}) are loaded for each virion produced.

We can now evaluate the survivorship function of a hypothetical cohort of T cells, as a function of the time spent exposed to the hazard of being detected by the immune system. The algorithm implements the survivorship function given by an integral equation (2.32):

- State values for N_r,N_s and the original number of class I-MHC molecules.
- State a value for N_{pep}.
- Select a functional form for viral production, such as those given by (2.52) and (2.53).
- Discretize a plausible life span for an newly infected T cell. For each age interval of this discretization calculate the number of virions produced, and increase the amount of loaded class I-MHC molecules according to N_{pep}. The probability of finding one loaded class I-MHC molecule in a particular age interval, $pN_r(a)$, is calculated by dividing the current number of loaded molecules by the original number of unloaded molecules. The probability that immune system will find N_r out of N_s loaded class I-MHC molecules is:

$$l(a) \quad = \quad \text{Bino}\left(N_r, N_s, pN_r(a)\right) \tag{2.54}$$

where Bino $(N_r, N_s, pN_r(a))$ is the cumulative binomial distribution, for N_r successes out of N_s trials, given that the probability of one success is $pN_r(a)$.

A range of plausible model parameters leads to a survivorship function with a shape depicted in the Figures 2.4 and 2.5. The first figure assumes a delayed exponential type function for viral production. In the second figure a Hill type function with no delay is assumed. Note that apart from CTL mediated lysis, there is also a 'constant background' threat experienced by infected T cells. A constant increase/decrease in mortality over all ages, will however not dramatically affect the shape of a lifetime distribution (Appendix E). The argument that mechanisms at binding sites of infected T cells determine the shape of lifetime distributions of these cells, may therefore be extended to an environment of slowly varying immune system response.

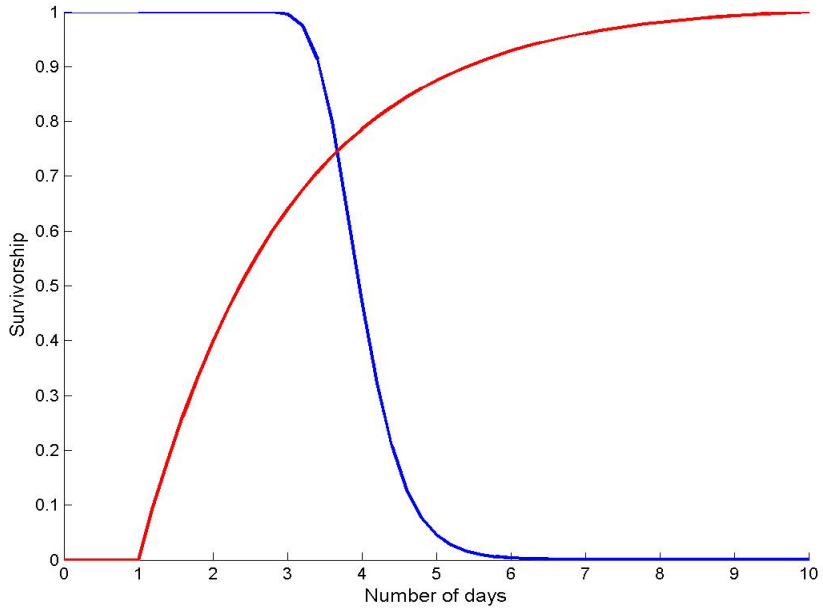

Fig. 2.4: The blue curve shows a decline in the immune system escaping ability of an infected T cell, due to increasing class I-MHC display. Normalized viral production is depicted in red. An exponential viral production function, delayed by one day, is assumed.

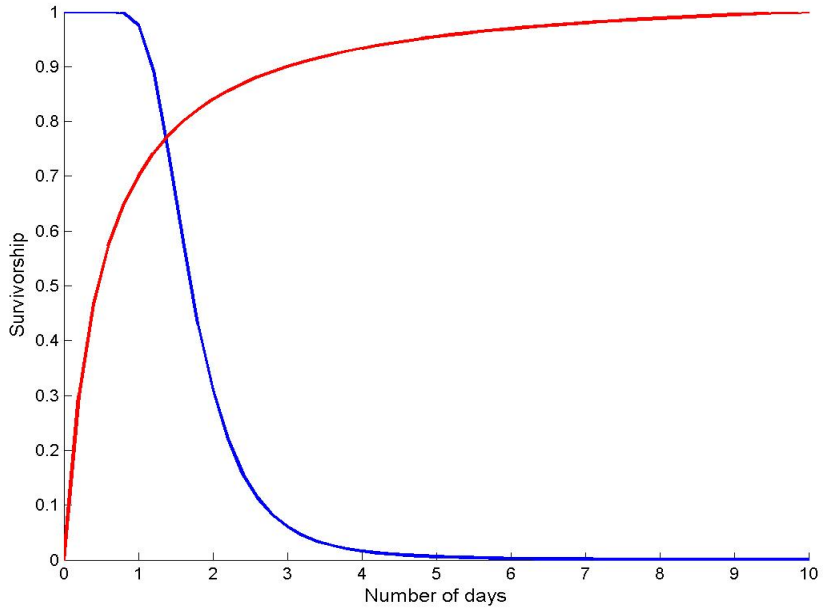

Fig. 2.5: The blue curve shows a decline in the immune system escaping ability of an infect T cell, due increasing class I-MHC display. Normalized viral production is depicted by means of the red curve.

The importance of this discussion seems to be that physiological insights and knowledge of binding mechanisms, can lead to constraints on the kinds of age dependence that parameters in age-structured models will exhibit. In the present example, these constraints arise entirely from the combinatorics of risk exposure at surface binding sites. A range of plausible model parameters leads to step-function-like survivorship rates; i.e. infected cells will go about unnoticed for a characteristic time, after which their guarantee for survival expires rapidly. The model derived in this section is by no means exact, but could be improved and adapted to incorporate additional physiology that determine the survivorship of infected cells.

2.6 A numerical tool for exploration of structured viral dynamics models

A simple model expressing survivorship of infected cells as a function of viral production and immune system response, has been discussed in Section 2.5. The purpose of this section is to build these survivorship functions into an age structured viral dynamics model, to solve this model numerically, and to discuss solutions in the context of the physiological constraints mentioned above.

Section 2.3 introduced an age structured model for populations of infected cells:

$$\frac{\partial P_{1,2}(a,t)}{\partial t} = -\frac{\partial P_{1,2}(a,t)}{\partial a} - \mu_P(a)\, P_{1,2}(a,t) \tag{2.55}$$

with initial conditions $P_{1,2}(a,0)$ and boundary conditions for $P_{1,2}(0,t)$. The main difficulty with a numerical implementation of equation (2.55) is the calculation of a continuous 'flow' of infected cells along the age axis. The EBT (Escalator Boxcar Train) method, introduced by [11, 9], is a powerful technique for studying the dynamical behavior of physiologically structured populations. The algorithm is built around simple rules designed to keep track of cohorts of population members: 1) the population is firstly divided into a collection of cohorts, in a manner permitting characterization of these cohorts by the number and total age of individuals it contains, 2) cohorts are tracked through time, subject to the rule that individuals stay in their cohorts unless removed by death, and 3) new cohorts are created through reproduction only. This is in contrast with the simplified model for gamma lifetime distribution, introduced in Section 2.4, where a continuous flow of individuals occurs across stage boundaries. Both approaches are built on the observation [9], that aging is a movement towards increased development. Consider a modification to the 2-strain model of equations (2.45)-(2.49), where the EBT algorithm is used to keep track of cohorts of infected cells:

$$\frac{dT(t)}{dt} = S_T(t) - T(t)\sum_{i=1}^{N_s} k_i V_i(t) - \mu_T T(t) \tag{2.56}$$

$$\frac{dP_{i,j}(t)}{dt} = \begin{cases} P_{i,j}(0,t) - \mu'_{P_i}(t_b)A_{i,j}(0,t) - \mu_{P_i}(t_b)P_{i,j}(t) & : \quad t_{j\Delta a} \\ -\mu_{P_i}(\overline{a_{i,j}},t)P_{i,j}(t) & : \quad \text{otherwise} \end{cases} \tag{2.57}$$

$$\frac{dA_{i,j}(t)}{dt} = \begin{cases} (t-t_b)P_{i,j}(0,t) + P_{i,j}(0,t) - \mu_{P_i}(t_b)A_{i,j}(t) & : \quad t_{j\Delta a} \\ P_{i,j}(t) - \mu_{P_i}(\overline{a_{i,j}},t)A_{i,j}(t) & : \quad \text{otherwise} \end{cases} \tag{2.58}$$

$$t_{j\Delta a} \to j\Delta a \le t \le (j+1)\Delta a \tag{2.59}$$

$$\overline{a_{i,j}}(t) = \frac{A_{i,j}(t)}{P_{i,j}(t)} \tag{2.60}$$

$$\frac{dV(t)}{dt} = N_{i,1}(\overline{a_{i,1}},t)P_{i,1}(t) + \ldots + N_{i,n}(\overline{a_{i,n}},t)P_{i,n}(t) - \mu_{V_j}V_j(t) \tag{2.61}$$

The origin of this algorithm, as well as factors that contribute to its accuracy, are discussed in Appendix D. It is demonstrated that the discretization scheme described by equations (2.56)- (2.61), becomes exact when the age interval is divided into intervals over which the mortality function μ_{P_i} is piecewise linear.

The following remarks are aimed at highlighting some of the features of the algorithm, originally designed to model a single population with physiological structure (size,mass,age, and so on), by means of its specific application to model *interacting* populations with *age* structure. These remarks describe the terms in the equations above and also highlight some useful implementation details. The algorithm was implemented from scratch in $Matlab$ for the present work.

- $P_{i,j}$, where $i = 1, 2$ and $j = 1, \ldots, n$, refer to the total count of individuals in n cohorts distinguished for the wild type $P_{1,j}$ and the drug resistant strain $P_{2,j}$ respectively. These cohorts are recruited during a time interval of $\Delta a = \frac{a_{max}}{n}$ each. It is assumed that individuals cells do not live longer than age a_{max}.

- Equations (2.58)-(2.60) are 'bookkeeping' equations for the total ($A_{i,j}$) age and the average age ($\overline{a_{i,j}}$) in each cohort. The term

$$P_{i,j}(0,t) \quad = \quad k_i V_i(t) T_i(t) + T(t) \sum_{j \neq 1}^{N} \varepsilon_{ij} k_i V_i(t)$$

  in equations (2.57) and (2.58) refers to 'newborn' infected cells that are recruited due to new infections of healthy cells. The conditional statement of (2.57) and (2.58) emphasizes that one boundary or birth cohort is distinguished into which all births are recruited. The ODE's for tracking the birth cohort are different from the ODE's tracking any other cohort. Notice that the average age $\overline{a_{i,j}}$ is used as a representative statistic for general cohorts (equation (2.57)), but that the boundary cohort (equation (2.58)) uses the time of birth t_b. The reason why the average age of the boundary cohort is considered to be an unreliable statistic, is that the average age is not well defined when $P_{i,j}(t)$ is zero in equation (2.60). This could happen during the initial stage of integration, as the boundary cohort is initially empty. Equation (2.58), tracking the total age of a cohort, describes the rate of change of the total age ($A_{i,j}$) due to aging and mortality [3]. Thus a population of infected cells can be viewed as moving along an escalator, hence the name of the algorithm.

- $\mu_{P_i}(\overline{a_{i,j}}, t)$ is the rate at which strain cohort $P_{i,j}$ is cleared from the system, when the average age of the cohort is $\overline{a_{i,j}}$.

- $N_{i,j}(\overline{a_{i,j}}, t)$ represents the rate at which virions are produced by a cohort $P_{i,j}$ with an average age of $\overline{a_{i,j}}$.

- The model of equations (2.56) - (2.61) can be implemented in a simple way according to the following scheme. Any integration algorithm, providing an interface to specify the derivative of state variables (i.e. the right hand sides of equations (2.56) to (2.61)), and a time range over which to integrate, will do. Equations (2.56) and (2.61) are straightforward to integrate, by 'passing' the

derivatives $\frac{dT(t)}{dt}$ and $\frac{dV(t)}{dt}$ to the integration routine. The latter equation will require the average age $\overline{a_{i,j}}$ of each cohort to be passed the routine calculating derivatives. The average age is also required for the calculation of derivatives $\frac{dP_{i,j}(t)}{dt}$ and $\frac{dA_{i,j}(t)}{dt}$, and hence the latest estimate for $\overline{a_{i,j}}$ must be continuously available. Tracking the state variables defined by equations (2.56) and (2.58) is slightly more tricky but manageable when the integration of the system of ODE's is stopped at every Δa interval. Each $P_{i,j}$ and $A_{i,j}$ cohort is moved along the age axis, i.e. $P_{i,j} = P_{i,j-1}$ for $j = 2, \ldots, n$, while new productively infected cells are recruited into cohort $P_{i,1}$. With this scheme, a fixed number of ODE's are maintained, which simplifies implementation.

- *Matlab* supports the parsing of string expressions into numerical values. This allows for great flexibility in implementing this algorithm, as we briefly explain. 'Matlab' provides a wide range of ODE algorithms and they all require as input: 1) the time derivatives of the system, 2) the integration range, and 3) parameters to optimize integration. These are two main aspects of the implementation. The time derivatives are first written as string expressions, containing terms for state and other variables. These string expressions are then parsed using the current values of all variables used in the expression, resulting in a vector containing the numerical values of the time derivatives, one value for each ODE in equations (2.56) and (2.58). We have already mentioned that the integration algorithm is stopped at each Δa interval in order to 'move the cohorts along the escalator'. This is the second aspect of the implementation and determines the integration range. The flexibility of this implementation is due to the fact that a new system of ODE's is generated for each choice of the mortality functions ($\mu_{P_1}(a)$ and $\mu_{P_2}(a)$) and discretization of the age interval.

- The EBT algorithm is accurate only when individuals in each cohort $P_{i,j}$ are accurately described by their average age $\overline{a_{i,j}}$. High accuracy can be achieved 1) by using a model with high resolution (small Δa), and 2) when mortality can be approximated by a piecewise linear function. The first option comes with increased computational effort, a typical tradeoff between resolution and accuracy. The second option allows for the possibility of an accurate low resolution model, provided that mortality is piecewise linear.

2.7 Exploring structured viral dynamics models

Some effort has been invested in exploring age structured models in the context of known clinical features of viral dynamical systems. The dynamics of an age structured model during primary infection, for example, have been studied by Nelson et al [20]. They noted that it is possible to find production schedules that fit primary infection studied by Stafford et al [24]. Another important issue regarding HIV progression, is the extent to which viral loads are reduced due to immune system

responses and viral cytopathicity respectively. A study by Nowak and May [6] interprets the consistency of viral decay slopes for patients undergoing drug therapy as evidence for the cytopathicity of the HIV virus. Simple mathematical models show that a fast acting immune system response can eliminate a significant fraction of the productively infected cell population, and thereby reduce viral production, without having much of an effect on the average life of infected cells.

This kind of analysis often highlights some of the discrepancies between mathematical models and clinical results. The drawback of this pursuit is that merely meeting test case standards does not necessarily produce models that could advance physiological and clinical insights, as much as simply following them. It would clearly be desirable for mathematical models to predict and explain dynamical behavior that would otherwise remain hidden. The purpose of this section is to explore the impact that our assumptions about the age distributions of infected cells have on strain competition.

Fitness implications of the maternity (viral production) schedule

The fundamental relationship between the intrinsic growth rate and the maternity schedule of a single self-generating population have been summarized in Appendix E. It follows from the observation that the characteristic equation (E.18) is the Laplace transform of the maternity schedule $m(a)l(0,a)$:

$$g(r) \quad = \quad \int_0^\infty e^{-ra} m(a) l(a)\, da \tag{2.62}$$

Normalizing the maternity schedule with respect to R_0:

$$\frac{g(r)}{R_0} = \int_0^\infty e^{-ra} \frac{m(a)l(a)}{R_0}\, da \tag{2.63}$$

and taking the logarithm of $\frac{g(r)}{R_0}$ generates an expression for R_0 in terms of r, μ and σ:

$$-\log(R_0) = -r\mu + \frac{r^2\sigma^2}{2!} - \ldots \tag{2.64}$$

where,

- r is the intrinsic growth rate.
- R_0 is the reproductive number.
- μ and σ is the mean and variance of the maternity schedule respectively.
- Moments of higher order than σ^2 have been omitted.

Using equation (2.64), [2] performed a perturbation analysis, providing an expression (detailed in Appendix E) for comparing the intrinsic growth rates of populations with different R_0, μ and σ respectively (all other parameters being equal). The following summary captures the essential points of this useful result:

- The population with a larger reproductive value $R_0 = \int_0^\infty m_i(a) l_i(a)\, da$ will have a higher intrinsic growth rate.
- A population with a larger average age at which birth is given, will have a lower intrinsic growth rate. Convergence to the stable age distribution is more rapid when the average fertility is localized at young ages.
- A population with a more dispersed age at which birth is given, will have a higher intrinsic growth rate.

Consider the two strain age structured model of equations (2.33)-(2.33):

$$\frac{dT(t)}{dt} = S_T(t) - T(t)\left(\sum_{i=1}^{N_s} k_i V_i(t)\right) - \mu_T T(t) \tag{2.65}$$

$$\frac{\partial P_i(a,t)}{\partial t} = -\frac{\partial P_i(a,t)}{\partial a} - \mu_{P_i}(a) P_i \tag{2.66}$$

$$P_i(0,t) = f k_i V_i(t) T_i(t) + T(t)\left(\sum_{j\neq i}^{N_s} \varepsilon_{ji} k_j V_j(t)\right) \tag{2.67}$$

$$\frac{dV_i(t)}{dt} = \int_0^\infty m_i(a) P_i(a,t)\, da - \mu_{V_i} V_i(t) \tag{2.68}$$

where $N_s = 2$, $i = 1, 2$ and $j = 1, 2$.

The dynamical behavior of the two populations of infected cells P_1, P_2 are coupled. They mutate into each other and compete for the same healthy T cells. It is instructive to investigate the impact that two different maternity schedules would have on strain competition, even though the theoretical result mentioned above only holds for the idealization of a single self-generating population. Figure 2.6 depicts the hypothetical age-structures of two strains. Information in this figure is organized as follows:

- Survivorship functions $l(a)$ are shown in blue.
- Lifetime distributions $L(a)$ are shown in red.
- Maternity or production function $m(a)$ are shown in green.
- Age dependent mortality rates $\mu(a)$ are shown in black.
- The lifetime distributions are gamma distributions and they are not implicitly related to maternity functions $m(a)$. Both age structures are hypothetical and may not comply with physiological constraints.

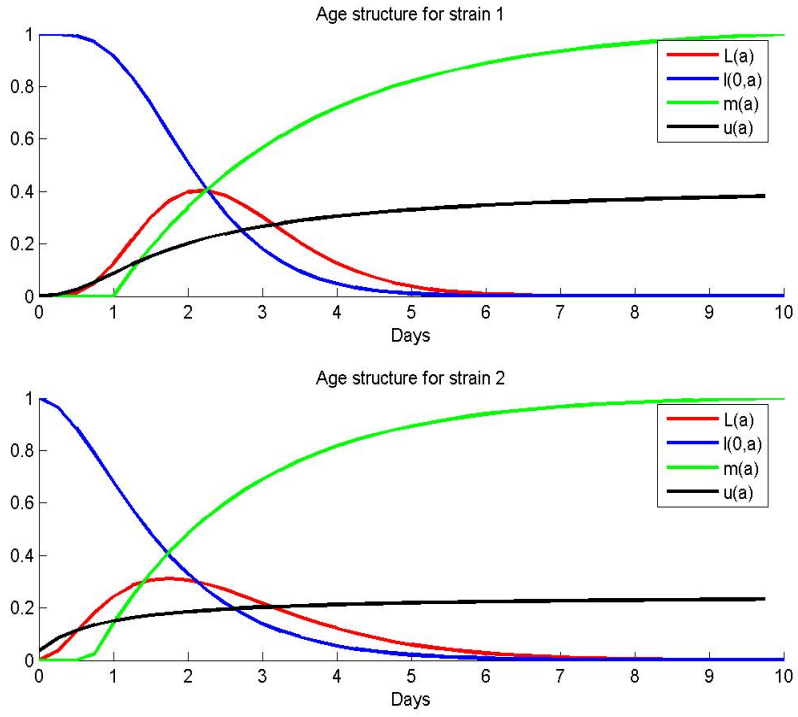

Fig. 2.6: Two viral strains with different hypothetical age structures. These age structures were generated by two slightly different sets of parameters for gamma lifetime distributions and exponential virion production functions. The values of these parameters are purely hypothetical. These age structures can be generated with the Matlab code provided in Appendix D.1

Recall from Section 2.3 that burst size N has been defined as:

$$N = \int_0^\infty m(a)l(a)\,da \tag{2.69}$$

The maternity distribution is obtained by normalizing the maternity schedule with respect to burst size:

$$M(a) = \int_0^\infty \frac{m(a)l(a)}{N}\,da \tag{2.70}$$

and is displayed in Figure 2.7 for the two strains.

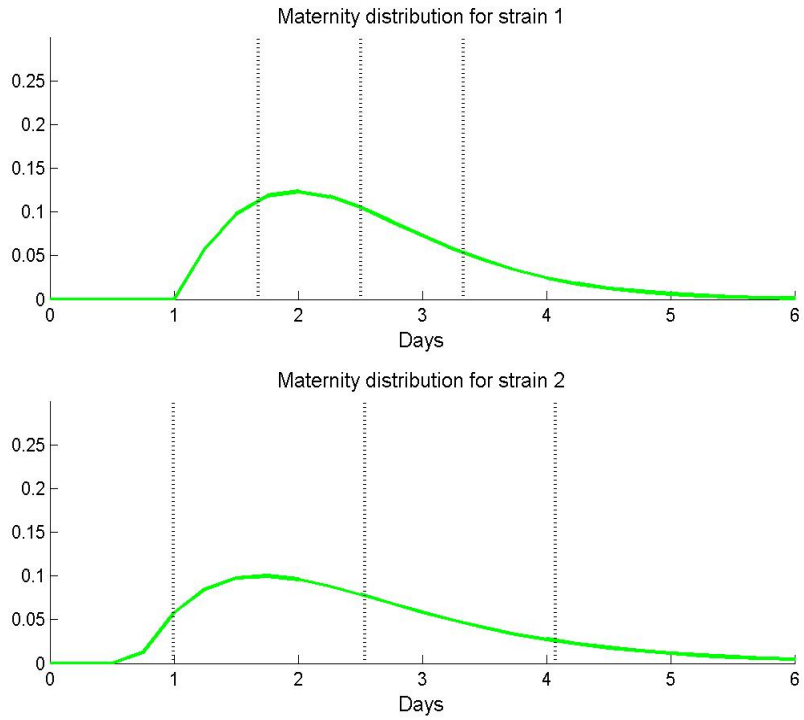

Fig. 2.7: Maternity distributions corresponding to the age structures depicted in Figure 2.6.

Notice that these maternity distributions:

- Have the same mean.
- They were also chosen to have the same burst size, namely 500 virions born to infected cells of either type.
- They differ however in the variance of the age at which new virions are produced: strain 2 has a higher variance in the age at giving birth to new virions.

We can expect strain 2 to dominate strain 1 as Figures 2.8 and 2.9 show.

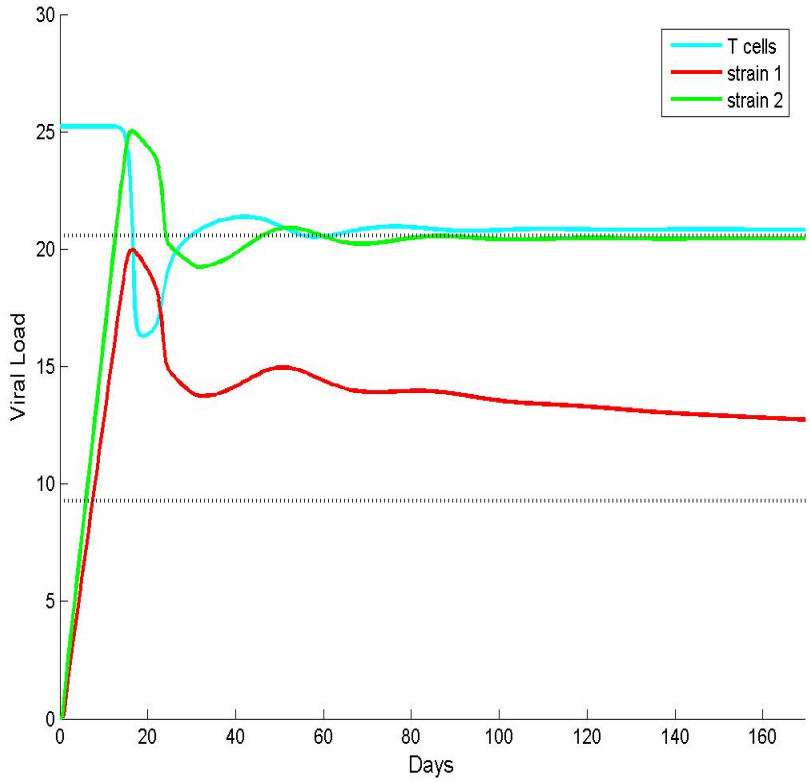

Fig. 2.8: Illustrating the dominance of a strain with a higher variance in age of giving birth. Both strains have mean age at birth of 2.6, i.e. infected cells give birth at an average age of 2.6 days. Both strains have a burst size of 500 virions and start with an initial viral count of 10. Strain 2 outcompetes strain 1 due to the fact that it has a greater variance with respect to the age at giving birth to new virions.

It can be seen from these two examples that fitness advantages implied by the maternity schedules of idealized self-generating populations, can still be seen when populations are coupled by competition and mutation. The fitness arguments for single self-generating populations derive from assumptions about their stable distribution. The question arises as to how much time can we expect populations to spend following their respective stable age distributions and asymptotic growth rates, and if stable age considerations are relevant for competing populations with a stable equilibrium. The following argument can be used to construct a scenario in which the stable age distribution and growth rates are well defined [32]: One could argue that the initial infection starts from a very small initial viral population. The initial target population is very large and remains roughly constant, at least for a period long enough for both populations to have reached their respective asymptotic growth rates and stable age distributions. The intrinsic growth rates of the two strains can

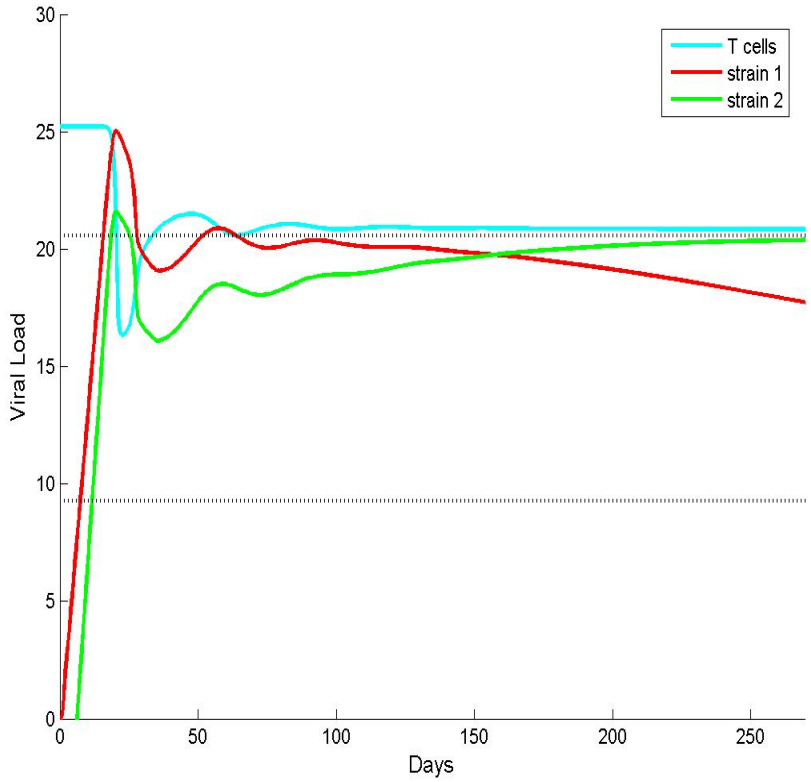

Fig. 2.9: The model is essentially the same as the model depicted in Figure 2.7. However, in this model strain 2 starts with an initial count of zero, and is created from mutating strain 1. The slight fitness advantage of strain 2 allows it to replace and dominate strain 1 eventually, even though strain 1 has 'won the race' to dominance during primary infection.

be equated with the slope of the initially linear parts on the logarithmic plots above. Strain 2, the fitter strain, has a steeper slope and therefore a greater intrinsic growth rate.

This digression shows that results from mathematical population dynamics, combined with numerical tools, can advance understanding of viral evolution. While it seems intuitively obvious that the strain with the greatest burst size will dominate 'less fit' strains, it is less obvious that maternity distributions, given by equation (2.70) will have such an important impact.

Modelling strain turnover dynamics with simplified lifetime distributions

In the previous section we have made use of an approximation scheme (i.e the EBT algorithm) for the system PDE's governing age-structured populations to implement

arbitrary age and maternity distributions. Certain aspects of strain competition and turnover dynamics can also be modelled using the simplified-lifetime-distribution approach developed in Section 2.4. We now investigate the consequences of assumed lifetime distributions for strain turnover rates. Consider again the two-strain model of Section 2.4 that uses the 'linear chain trick' to model gamma-like lifetime distributions. The model is put in equilibrium to avoid the dynamics of primary infection when studying the time required for a fitter mutant to replace the wild-type strain. The system is time-evolved starting from the equilibrium values for T, P_1 and V_1. The mutant is now 'created from rare mutation' by means the initial conditions $P_2 = 1$ and $V_2 = 0$, indicating that one cell, productively infected by the rare mutant, is introduced to the system. This scenario can only be physically possible if the wild-type and mutant have not been mutating into each other. The mutant produces $N_2 = 600$ virions compared to $N_1 = 500$ virions produced by the wild type. Its fitness advantage could be the result of the immune system being unprimed for the specific peptides (see Section 2.5) produced by the rare mutant. The time required by the mutant to replace the wild type is depicted in Figure 2.10 below. The lifetime distributions of cells infected by the two competing strains are given a shorter tail in each subplot, while keeping the mean age fixed. These lifetime distributions are depicted in Figure 2.3. It can be seen that the time required for replacement by an emerging strain, depends on the lifetime distribution of both competing strains, and ultimately on the rate at which existing cohorts are replaced by newly created cohorts. A short-tailed lifetime distribution compounds a fitness advantage, by allowing it to be used more often in the fitness competition that is being played out all the time. Notice that 'time to replacement' can be divided into three categories: 1) a distinctly longer time required by infected cells following exponential lifetime distributions, 2) a much shorter time required when these lifetime distributions are short-tailed, and 3) the time required for replacement tends to level off beyond a certain shape, which corresponds to roughly after 9 stages in the Figure 2.10 below. In Chapter 5 we infer the approximate shape of the lifetime distribution by assigning a probability to each of these categories. Knowledge of the precise value of 'the number of stages' in the life cycle of infected cells is not as important as simply knowing if the distribution is shorter-tailed than the usually assumed long-tailed exponential lifetime distribution.

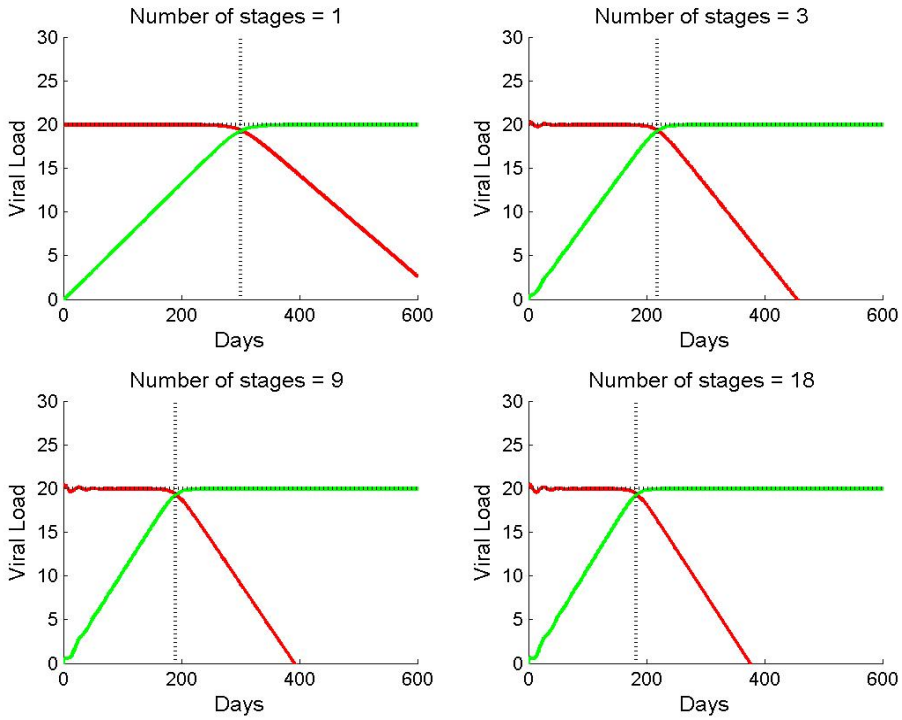

Fig. 2.10: Demonstrating a shorter time to replacement when a mutant follows a gamma-like lifetime distribution. The vertical line indicates the time required for replacement.

It is again demonstrated that the shape of a lifetime distribution determines a fitness differential beyond those that could be explained by the 'burst size' and mean lifetime of an infected cell. A mutant can emerge: 1) through a rare mutation event, 2) an immune system escaping strain or 3) as a resistant strain during anti-viral treatment. The shape of the lifetime distribution is a mechanism that determines the time to replacement in both cases.

A closer look at strain competition using scaling arguments

Structured population dynamics aim to find ways of structuring individual detail into models for dynamics at a population level. Models for HIV dynamics govern the interaction between billions of target cells and virions, and the method of averaging over individual detail is often used. This leads, for example, to the use of: 1) μ_P - the average lifetime of infected cells, 2) N - the average number of virions produced by an infected cells, and so on. This chapter has also focused on building details of the lifetime distribution in appropriate models. To this end, we have used staged ODE's to model gamma lifetime distributions and also PDE's to model lifetime

distributions with arbitrary shape.

This question could also be asked in a different and in a sense opposite way: what individual details can we be ignorant about? *Scaling arguments* are often used to help answer such questions, and we next investigate a simple scaling argument in the context of strain competition. In the previous section we have demonstrated the shorter time to replacement when a rare mutant follows a short-tailed lifetime distribution, and from Figure 2.10 it appears that strain competition can be completely described by exponential growth/decline of the two strains. This is concluded from the linearity of the growth curves on the log scale used in the figure. To investigate this phenomena in more detail we make use of the EBT algorithm, which allows us to model lifetime distributions of arbitrary shape. Another motivation is to ensure that observations made thus far do not depend on the use of gamma-lifetime distributions and staged ODE's. Using an lifetime distribution with shape given by the hypothetical age structure of the wild-type strain depicted in Figure 2.6, where the rare mutant is given a fitness advantage by increasing its production of virions by a factor of two across all ages, leads to a similar dynamical situation:

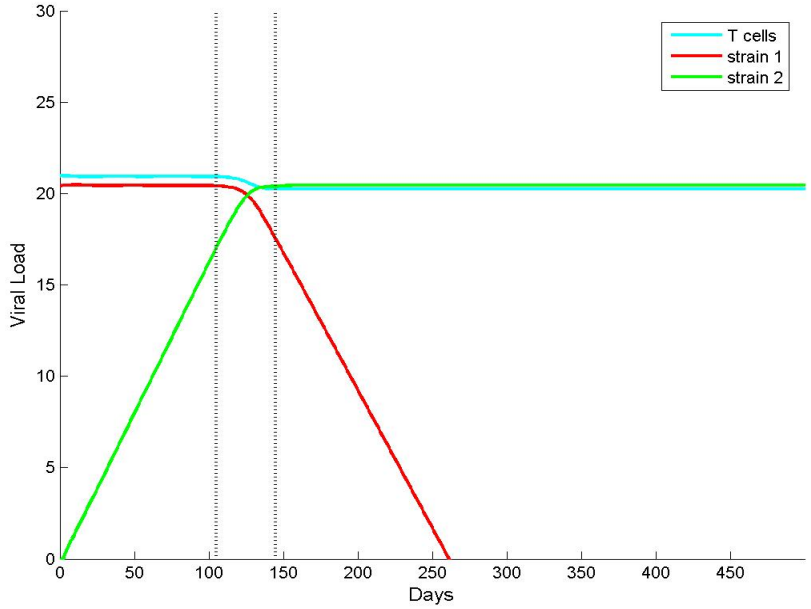

Fig. 2.11: Strain competition is completely described by the slope of growth curves during 'early infection'. This is the period when a growing infection is 'small in numbers' compared to other populations it interacts with. Strain 2, for example, remains small in numbers compared to the total count if Strain 2 and T cell populations, until a region near the cross over point is reached.

The equilibrium condition for any two strain model with differential fitness and no cross-mutation, is the eventual extinction of all but the fittest strain. This is the asymptotic behavior of the system. But there is also an intermediate and substantially large time region of linear behavior in the two strain model depicted above, namely the region of 'early infection'. We can make the observation that all environmental and individual detail is summarized into a single parameter, namely the intrinsic growth rate of a curve during 'initial infection'. A model consisting of two growth rates provides a complete description of strain replacement dynamics, at all times, except for a narrow time window when interacting populations coexist in comparable numbers.

2.8 Summary

This chapter has investigated mathematical models for HIV-dynamics. These models range from simple compartmental models, modelling simple virus-immune-system dynamics, to complex models that include general structure for populations of infected cells. Models for populations with physiological structure have been the focus of this research project, as simple compartmental models do not predict the rapid emergence of resistance during anti-viral treatment, at least not for plausible parameter values.

It was shown that simplified age distributions, such as gamma distributions, can be modelled quite easily by introducing suitable categories of stages traversed by infected cells. Resistance emerges more rapidly when the lifetimes of infected cells follow delta-like distributions, and the shape of these distributions were controlled by varying the number of 'stages' that an infected cell passes through. Using a simple algorithm that counts the number of viral peptides loaded on MHC complexes, it was demonstrated that infected cells will evade the immune system for a characteristic time, after which the probability of escaping immune system surveillance rapidly diminishes. This argument may justify the use of short-tailed lifetime distributions.

The seminal work of Ho et al [23] elucidated the rapid turnover rate of populations of infected cells. The mean of an assumed exponential lifetime distribution was inferred by fitting clinical data, obtained during therapy, to a simple model for HIV dynamics. Few other clinical insights have been directly obtained from mathematical models. The EBT algorithm was used to (numerically) explore the dynamics of two competing strains with hypothetical age structures. It was demonstrated that viral dynamics depend in detail on the whole maternity distribution. Mathematical models have focused mainly on the mean of the lifetime distribution. It may be instructive to investigate whether studying structures at the binding sites, together with other physiological facts of infected cells, may lead to a better understanding of infected cell lifetime distributions and associated viral production schedules.

This chapter also investigated methods for structuring some viral-life cycle details into mathematical models. The last section asked the 'opposite' question: which details can we be ignorant about. An argument exploiting the relative constancy of

state variables, demonstrates that all environmental and intrinsic fitness variables are summarized into one variable. This variable was found to be the slope of exponential growth curves, giving an accurate description of viral dynamics during a period of 'early infection'. This observation has a general context: Can mathematical analysis provide practical methods that can be used both to find relevant variables, and the time scale over which they are relevant?

Chapter 3

Measuring small quantities of DNA

The most frequently used technique for the amplification and quantification of specific sequences of deoxyribonucleic acid (DNA) molecules, is generally known as PCR (for Polymerase Chain Reaction). This technique implements the natural process of enzymatic DNA replication. During each temperature cycle, primers attach to DNA regions for which they are (designed to be) complementary. The primer is then extended to duplicate the target sequence. The number of target DNA molecules is ideally doubled in each cycle, leading to an exponential replication process with a theoretical efficiency of 2. This level of efficiency is usually not attained in practice due to a number of contributing factors: 1) nonspecific target amplification, 2) enzyme degradation due to short enzyme lifetime, and 3) product inhibition, to name but a few.

A 'Real Time' PCR (RT-PCR) machine performs a measurement (by fluorescence - to be discussed in detail later) of the amount of DNA present at each cycle. This measurement is dominated by background noise during the initial cycles because the amount of DNA is so small that it is essentially undetectable. The amount of DNA grows exponentially with each PCR cycle, and its measurement soon dominates the background signal. At this point the measurement can be used to estimate the unknown initial concentration of DNA. The more cycles of PCR required to reach a standard quantity of PCR product, the less there was present to begin with.

A picture of viral evolution and strain competition can be formed by measuring how the viral loads (or their relative fractions) of competing strains change over time. The purpose of this chapter is to study the PCR process for DNA synthesis in some detail, to estimate the relative concentrations of viral DNA sequences, and to estimate the uncertainties in these estimates due to the imperfect PCR and measurement processes.

A *Matlab* toolbox was developed to process PCR data. These routines use a simple mathematical formulation for the branching probability of each DNA sequence. Although the model is idealized, it is convenient and accurate in explaining the exponential growth phase of the PCR process. The toolbox implements a few standard methods and introduces two unique methods of processing PCR data. The

first technique estimates the background signal in the measurement process and removes it from the data. The second method is a correction to subtle effects introduced into PCR data by non-specific (unintended) amplification. This correction is applied to estimates of the magnitudes of relative fractions of viral subpopulations. The correction derives from calibration PCR runs, where primers are used to amplify plasmids (commercially manufactured DNA sequences) that they are not specifically designed to amplify. Modelling these mispriming, or spurious amplification effects, shows that these errors are compounded by each PCR cycle. A conclusion of this analysis is that a small population of DNA may have been unintentionally created from other DNA populations, and that caution should be applied when reporting that they exist *in vivo*. Accurate quantification of small viral subpopulations is a crucial component of this project, as it provides us with data that could be used to calibrate the mathematical models developed in Chapter 2. An alternative PCR method that does not compound mispriming errors, is discussed in the summary of this chapter.

3.1 Genetics: basic concepts

The purpose of this section is to provide a simple introduction to some of the concepts and terminology of genetics and its molecular structure, regarded as relevant to study PCR. Compact, but informative, introductions can be found in [33, 34].

Genetic instructions are carried in deoxyribonucleic acid (DNA). This is a nucleic acid consisting of a string of covalently-bound nucleotides. Encoding is achieved with four nucleotides (also called bases): adenine (A), thymine (T), cytosine (C), and guanine (G). A and G are known as purines and C and T as pyrimidines. A 'pairs up' with T, and C with G, by means of hydrogen bonds, forming a double helix. A DNA sequence can be replicated by splitting (dissociating) the double strand. This provides two single strands that can be used as templates for synthesizing two identical copies of the original strands. Mutations, or errors of the copying process, occur when nucleotides are incorrectly copied, skipped or inserted.

The genetic code is made up of three letter words termed codons, such as ACT, CAG, TTT, and so on, up to 64 possibilities. A codon corresponds to a amino acid. This is a many to one relationship, as there are only 20 amino acids but 64 codons. Only some regions of the genome encode for proteins. Encoding regions are 'demarcated' by, and is situated between, 'stop' codons.

The chemical structure of ribonucleic acid (RNA) is very similar to that of DNA. RNA molecules have an additional hydroxyl group. Another difference is the use of uracil (U) instead of thymine (T). RNA is transcribed from a DNA template using RNA polymerase enzymes. Viral genetic material can be found in distinct compartments and can be related to distinct viral RNA activities:

- Viral RNA is first turned into DNA by means of the reverse transcriptase enzyme before it is integrated into the host cell genome by the integrase enzyme.

- One of the main functions of RNA is to translate genetic information from DNA into proteins. Messenger RNA (mRNA) is transported to ribosomes in the cytoplasm of the cell, where the information is translated into instructions for protein synthesis. Quantifying mRNA is called 'gene expression' or 'gene activity' analysis.

- Free virus, circulating in the body, is essentially a piece of viral RNA protected by a capsid.

3.2 The PCR process

The PCR system is a method for *in vitro* DNA replication and provides a quantitative method for detecting small quantities of DNA. The PCR process uses polymerase enzymes to replicate DNA. Attention is focused on a particular region of the DNA molecule, selected for amplification by means of *primers*. These are short sequences of synthetic DNA that are complementary to the sequence to be amplified.

In this project we analyzed data from a PCR experiment which was designed to quantify the distribution of the nucleotides A,C,G, and T at codon 103 at the *env* gene. This codon is of interest because 103_A and 103_G encode for a drug (Nevirapine) sensitive genotype, while 103_C and 103_T encode proteins for a resistant genotype. Primers 103_A,103_C,103_G and 103_T will refer to primers targeting viral strains 103_A,103_C,103_G and 103_T respectively.

An initial small amount of DNA (template) is placed in a well, together with primers, polymerase enzymes, and other biochemical components. The PCR cycle then proceeds through the following steps:

- *Denaturation:* The temperature is raised to $94°C$, for approximately one minute. At this temperature double stranded DNA opens into single strands. Enzymes used in this process (usually the Taq polymerase enzyme) are designed to survive this high temperature. Note that enzymatic reactions will stop at this high temperature, resulting in incomplete extensions and populations of synthesized DNA with different lengths.

- *Annealing:* The temperature is lowered to $54°C$, allowing primers to attach to the single stranded template. Stable attachments allow the polymerase enzymes to attach (hybridize) to the small section of double stranded DNA (formed by the primer and the template) and copying of the template can start.

- *Extension:* The temperature is again raised to $72°C$, the optimal temperature for extension of the template by the polymerase enzymes. Loose bonds, from primers not matching the template exactly, break and will therefore not be extended. This limits the amplification of non-specific primer-template matches.

These steps are repeated for approximately 40 cycles by a thermocycler, programmed to cycle between the denaturation, annealing, and extension temperatures. The temperatures mentioned above are specific to to a particular process and depend on physical properties of the nucleic acids participating in the reaction.

3.3 Measuring PCR product

The PCR experiment used to generate the data of this project includes one 'reporter dye' in the biochemical mix of each well. This dye interchelates non-specifically into the double stranded DNA generated during each PCR cycle. The dye will generate a characteristic emission profile only if interchelated, when light (from the UV part of the spectrum) is projected through the well. This fluorescence profile is monitored during the extension phase of each cycle. In practise, the data from the experiment is an electric signal from a detector, which has a linear response to fluorescence in only a part of its dynamical range. A passive dye is also added, and is part of the master mix which forms the basis of the biochemistry in each well. The passive dye does not interchelate into double stranded DNA and its emission profile is used to monitor fluctuations that result from PCR independent processes. Figures 3.1 and 3.2 show emission profiles for a passive (ROX) and a reporter dye (SYBR) respectively.

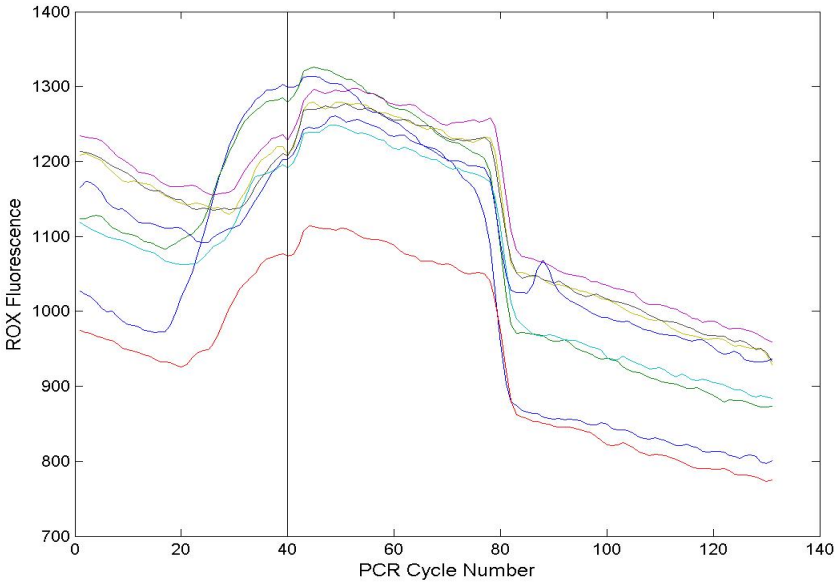

Fig. 3.1: Passive dye(ROX) fluorescence. The PCR process does not extend beyond the vertical line, although measuring of the degrading product continues.

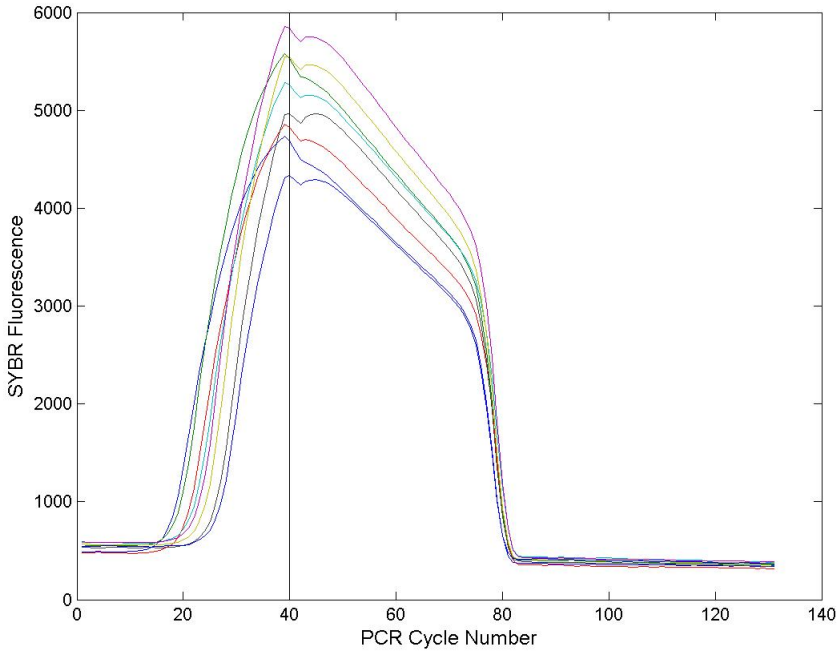

Fig. 3.2: Reporter dye(SYBR) fluorescence.

A number of properties of the PCR process are reflected in these amplification plots. The process degrades toward cycle 40 as it runs out of biochemical ingredients. Thermal cycling is stopped after 40 cycles, ending the PCR process. Measurement of the degrading product continues. The cycles of interest lie between 10 and 35, when the reporter dye rises sharply above the background signal. The cycle number at which a reference quantity of product (P_T) is attained is called the threshold cycle (c_t). The passive dye signal also increases in this phase of the process. The apparent reason for this that the ROX detection channel picks up the tail of the SYBR emission spectrum.

Reporter signal is normalized (and then called Rn) after division by the passive signal. This corrects the reporter signal for well-specific fluctuations, allowing inter well comparison of reporter signal data [35]. A further correction to the reporter signal is made by subtracting a background signal. This corrected reported signal is usually called 'Delta Rn' (ΔR_n) and refers to a normalized reported signal from which the background has been removed.

3.4 Estimating PCR amplification rate

The simplest mathematical model for the PCR process assumes that the probability (m) for a single DNA molecule to duplicate during each cycle, remains constant:

$$P(c) \quad = \quad P(0)m^c = P(0)(1 + m_{eff})^c \tag{3.1}$$

where,

- $P(0)$ is the initial concentration of template.
- c is the cycle number.
- m is the amplification rate of PCR product.
- m_{eff} is the efficiency of the process.

Detector signal processing, background subtraction and normalization

A formula for the detection of PCR product has the following form:

$$R_n(c) = A + F\left(P(c)\right) \tag{3.2}$$

- R_n is normalized fluorescence.
- F is a detector response function which has a region of linear response to fluorescing PCR product.
- A is the background signal to be estimated and removed from the reported signal.

The fluorescence signal of the first few PCR cycles is dominated by measurement noise when synthesized DNA is present in too low concentration for the signal to rise above this background. The literature [35] suggests that simply subtracting the average fluorescence from the first few cycles, will correct the signal for background noise. We make use of a slightly more nuanced method to estimate the background signal.

The discrete 'cycle number' variable can be replaced by a continuous time variable to make use of this relationship. Measuring the PCR reaction may then be written as:

$$R_n(c) \quad = \quad A + F\left(P(0)e^{\lambda c}\right) \tag{3.3}$$

Fluorescing PCR product dominates background noise in the reported measurement:

$$A \ll R_n(c_t) = F\left(P(0)m^{c_t}\right) \tag{3.4}$$

A theoretical curve (depicted in Figure 3.3) can be defined in this region, and its initial concentration can be calculated as follows

$$F\left(P(0)\right) \quad = \quad F\left(P_T e^{-\lambda c_t}\right) \tag{3.5}$$

and inserted into equation (3.3) to obtain the measured signal at at arbitrary cycle c:

$$F\left(P(c)\right) \approx A + F\left(P_T e^{\lambda(c-c_t)}\right) \tag{3.6}$$

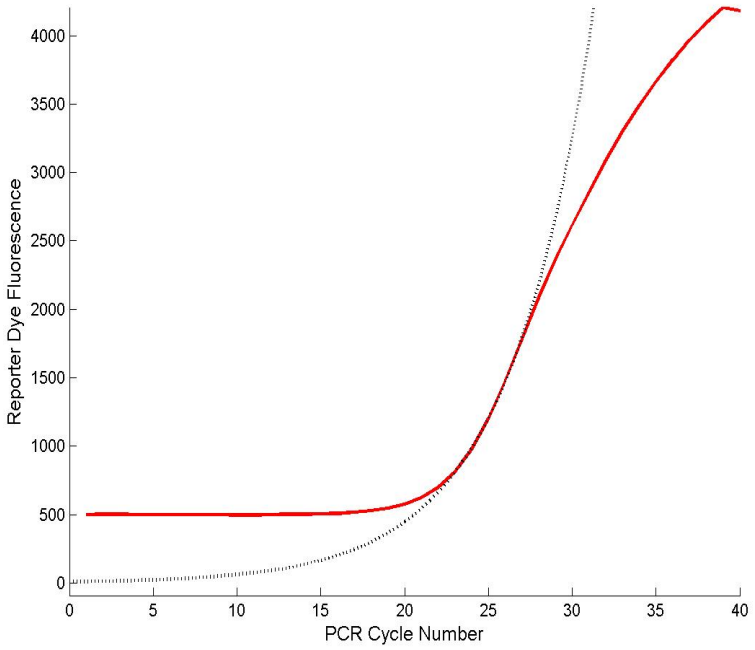

Fig. 3.3: Background signal (A), depicted with the apparent constant part of the red curve, dominates the fluorescence during the first few cycles. The theoretical curve, used to calculate an estimate for the latest cycle where measurement reports only the background signal, is depicted by the dotted black line.

We can now estimate the background signal as the average signal of the first c_A cycles, where c_A is the cycle number and where $F\left(P_T e^{\lambda(c-c_t)}\right)$ becomes more than 1 percent of what was previously called 'background signal'. This provides a more accurate description of the background signal, as it is based on the linear and most reliable part of the measured curve. All reporter signals now refer to a reporter signal that has been normalized with respect to the passive dye and corrected for background noise using this method. ΔR_n will refer to background-removed signals.

With the background removed, equation (3.3) reduces to:

$$F\left(P(c)\right) = F\left(P(0)e^{\lambda c}\right) \tag{3.7}$$

At this point we drop explicit reference to the detector response function F, and adopt

$$P(c) = P(0)e^{\lambda c} \tag{3.8}$$

as a working description of the measured PCR process. Several cycles of linear increase in fluorescence may be identified by plotting the normalized reporter signal on a log scale. We rewrite (3.8) as

$$ln\left(P(c)\right) \quad = \quad ln\left(P(0)\right) + \lambda c \tag{3.9}$$

This relationship can be used to find the most linear region in the exponential phase of the reaction and hence to estimate λ as the slope of this line. Several methods are available for estimating the amplification rate of the PCR process.

The 'Amplification Rate' method

In this method an exponential curve is fitted to the reporter signal. This is possible as the exponential growth curve is detectable for 3 to 4 cycles on a log-linear plot. Four dilution runs were performed in this study, using primers designed to target pure plasmids with an A, C, G or T at codon 103. The exponential phase can easily be identified as linear on a log plot. This method is illustrated with Figure 3.4, using the dilutions runs for plasmid A.

The exponential curves estimated from $log(\Delta R_n)$ plots are displayed in Figure 3.5. It can be seen that they are in good agreement with the fluorescence data in the exponential phase of the PCR process.

The standard curve method

This method relates the 'threshold cycle' required to attain a reference quantity of product across a set of wells where an initial (unknown quantity) of template is serially diluted. The dilution experiments in this study for Primers $103_A, 103_G$ and 103_T yielded accurate 'standard curves'. The dilution experiment for Primer 103_C was unsuccessful, necessitating a further investigation into the efficiency of the PCR reaction, in order to estimate the efficiency of Primer 103_C.

The 'threshold cycle' was estimated for each curve at the intersection of the curves depicted in Figure 3.5, with a user defined reference quantity (yellow line). The linearity of the c_t vs n relationship, where n is the two-fold dilution number, can be used to estimate the amplification rate. The leftmost curve, with the fastest rise time, corresponds to amplification of the initial (unknown amount of)

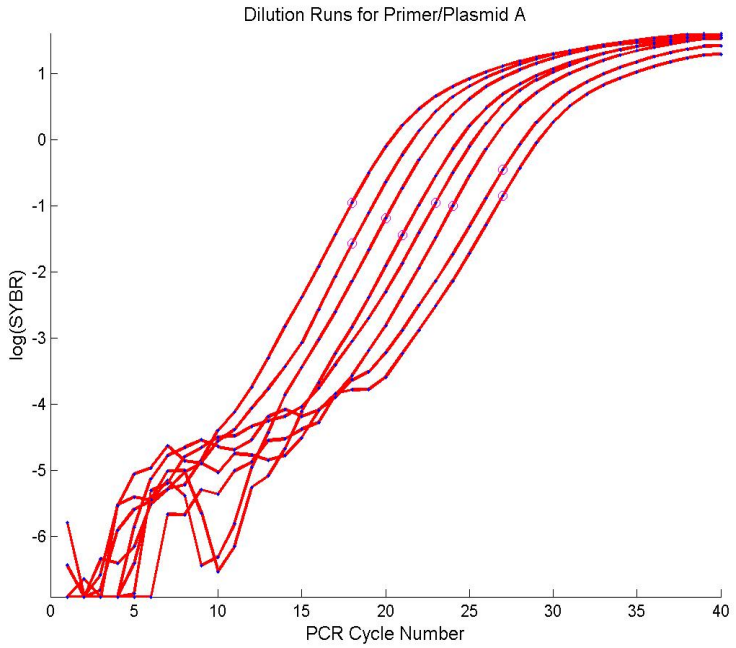

Fig. 3.4: Dilution runs for primer 103_A.

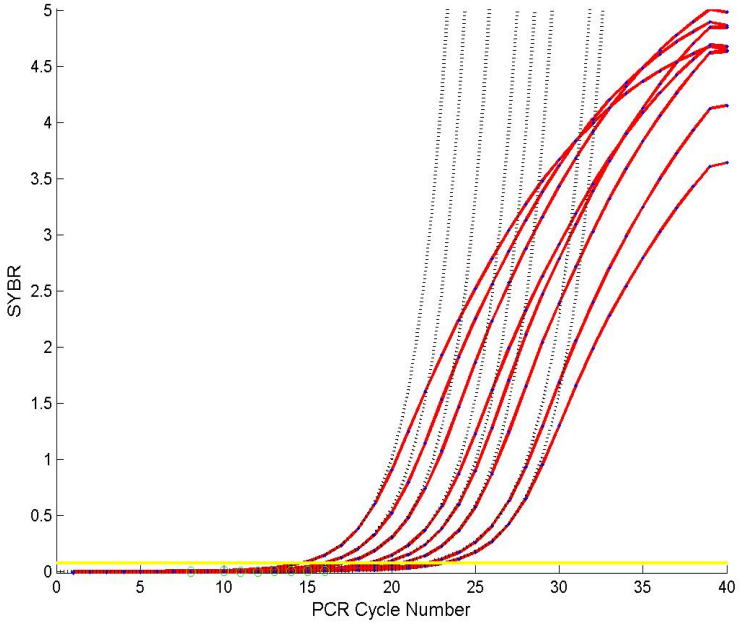

Fig. 3.5: Estimating amplification processes for primer 103_A.

template $P(0,0)$. If $P(0,n)$ represents the same template diluted n times, then $P(0,n) = \frac{P(0,0)}{2^n}$. The general relationship between threshold cycle and diluted initial concentration is given by:

$$P_T^{(n)} = P(0,n)m^{c_t}$$

and therefore:

$$P(0,0)m^{c_t} = P_T 2^n$$

It follows that the slope of the linear relationship between c_t and n equals $\frac{ln(2)}{ln(m)}$. If the slope of the line is given by S, then $m = e^{\frac{ln(2)}{S}}$, where S is estimated as the slope of the line depicted in Figure 3.6.

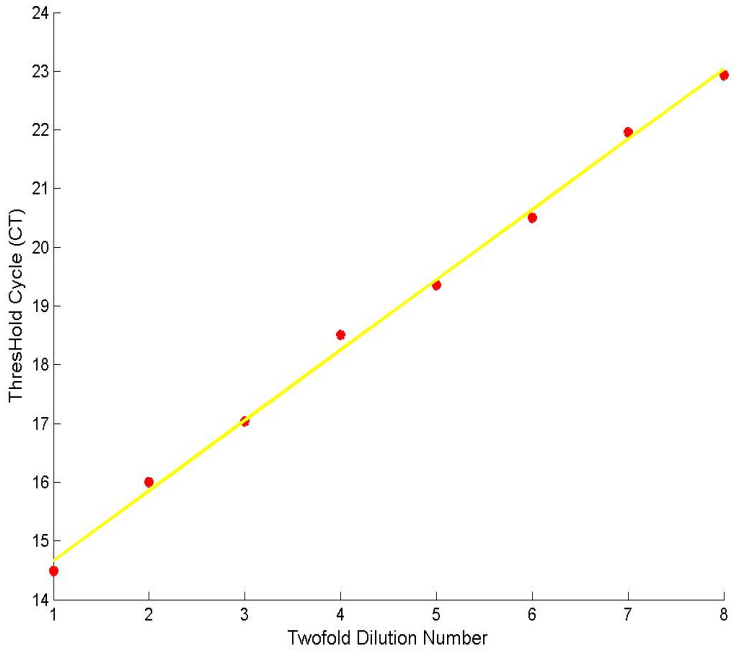

Fig. 3.6: Linearity of CT vs template dilution.

The 'Standard Curve' method gives a constant multiplication rate of 1.75 for the amplification rate of $103A$ template, whereas the 'Amplification Rate' method, i.e. that of curve fitting on individual wells, gives a set of values clustered around 1.65. This discrepancy between amplification rates estimated by the two methods does not have an obvious explanation.

The 'Standard Curve' method is based on the assumptions that: 1) the increasing number of cycles required to get a more diluted template to a threshold amount of product, is linearly related to the two-fold dilution number of the template, 2) the slope of this line, i.e. the amplification rate, is assumed to be unchanged for the duration of the number of cycles required to reach threshold, and 3) template dilution does not change the amplification rate. To the extent that these assumptions are accurate, the 'Standard curve method' gives an accurate estimate of PCR efficiency. This method seems to give an estimate of efficiency during the first few cycles, on the basis that the PCR process decreases in efficiency towards higher cycle numbers [36]. Note that two templates which differ by a two-fold dilution factor, will be of approximately the same concentration after (a bit more than) one cycle. The reaction will proceed identically, in an environment of perfect chemistry, as soon as the PCR product in the diluted well has 'caught up'. The linear c_t vs n relationship is therefore established during the (early) cycles required to get the diluted templates up to the concentration of less diluted templates. This takes roughly 10 cycles for the situation depicted in Figure 3.6.

Factors contributing to decline in efficiency

There are many factors contributing to the decline in efficiency and the better understood factors include [36]:

- The polymerase enzyme has an intrinsic half life of about 30 minutes, when subjected to the high temperatures of PCR thermal cycling. This half life is reached after about 30 PCR cycles. At this time half of the enzyme population is no longer viable for use in the process, but the enzyme loss is a process that starts before this time. This enzyme activity will contribute in a non trivial way, in terms of availability, viability, and the rate at which viable enzymes are used by the process.
- Within a few cycles after a process reaches threshold the reaction efficiency declines as ingredients are depleted, being rapidly used as a resource for replicating the exploding populations of synthesized DNA.
- As the population of synthesized DNA increases the relative availability of nucleotides decreases. This may lead to ineffective and incomplete primer extensions.
- Inhibition of all reactions with increasing concentration of inhibitors (byproducts of the reaction).
- Reduction in denaturation efficiency, as the denaturation temperature is typically not adjusted at each cycle to a temperature of optimal DNA separation.
- Optimal denaturation is only possible by understanding in detail the physics governing reactions in each well.

Estimating declining efficiency of PCR

The simple formula $P_T = P(0)e^{\lambda c_t}$ allows us to estimate the initial concentration in (arbitrary) units of fluorescence. However, simply using $P(0) = P_T e^{-\lambda c_t}$ seems naive in the light of the fact that the amplification λ appears to be a function which decreases with cycle number. This formula should be written as:

$$P_T = P(0)e^{\int_0^{c_t} \lambda(c)\, dc} \tag{3.10}$$

To estimate $\lambda(c)$ empirically, we pooled 220 amplification curves and looked at the relationship of the fitted amplification rates for each of the primers. The amplification rates are displayed in Figure 3.7, against cycle number, and a regression line is fitted to each data set.

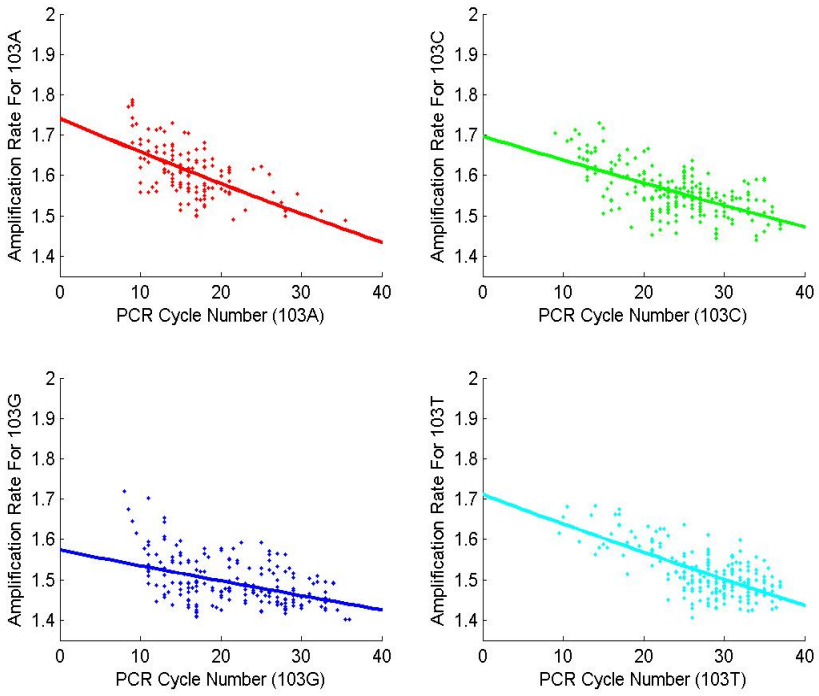

Fig. 3.7: Decline in amplification efficiency as a function of increasing cycle number.

It can be seen that the decline in efficiency can be observed to first order with a regression line. The regression line for the decline in efficiency of the PCR reaction with cycle number is listed below for each primer:

$$m_{103_A} = -0.0049c_t + 1.74 \quad (3.11)$$
$$m_{103_C} = -0.0035c_t + 1.69 \quad (3.12)$$
$$m_{103_G} = -0.0025c_t + 1.58 \quad (3.13)$$
$$m_{103_T} = -0.0044c_t + 1.71 \quad (3.14)$$
$$(3.15)$$

From the visible scatter of amplification rates it is clear that a regression line can only be a first order description. In reality the process is stochastic, and the amplification rate at each cycle can be viewed as a draw from an unknown distribution, possibly peaked around the regression lines. A stochastic formulation will however require much more information about the underlying PCR and detection processes.

This digression into PCR efficiency has raised questions that will remain unanswered until additional experiments and data sets become available. Primers are typically tailor-made for a particular target sequence, allowing rapid amplification, leaving a very narrow window for estimating PCR efficiency by the curve fitting methods. This PCR experiment was not designed to resolve PCR efficiency questions. The most important questions for the purposes of this project are: 1) do the primers have significantly different amplification efficiency, and 2) does Primer 103_C, designed to quantify the resistant mutant, behave much differently from Primer 103_A, designed to quantify the wild-type. Based on the amplification vs cycle number relationships depicted in Figure 3.7, and expressed in equations (3.11)- (3.14), the following assumptions are made:

- There is not enough information to warrant the use of different efficiency relationships $\lambda_A(c)$, $\lambda_C(c)$ and $\lambda_T(c)$.
- This decision is corroborated by the similar standard curves of Primers 103_A and 103_T, the standard curve for 103_C being absent.
- The standard curve for Primer 103_G along with the estimate for $\lambda_G(c)$ (equation (3.13)), suggests that Primer 103_G appears to be less efficient than the other primers. This is not expected to have a big implication for estimating initial fractions as 103_G viral RNA generally appear in very low concentrations.

3.5 Estimating initial concentrations of RNA

Initial concentrations calculated by equation (3.10) are represented by the red lines in Figure 3.8. The green curve represents exact two fold dilution starting with the initial concentration of the undiluted template, given by the intercept of the red curve. Estimated initial concentrations can be seen to be in close agreement with exact two fold dilution.

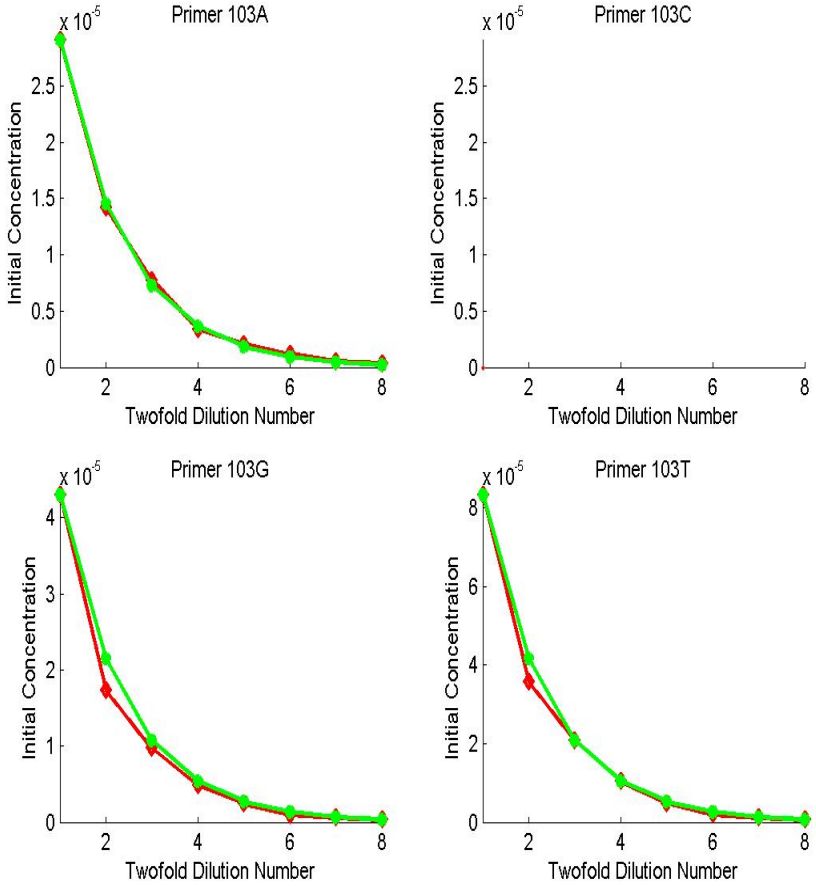

Fig. 3.8: Estimated initial concentration for dilutions runs of plasmid 103A,103G and 103T. Dilution runs for 103C were not successful.

Calibrating initial concentrations for measurable mispriming effects

It can be observed that primers generate spurious product from incorrect targets. What happens, approximately, is that the primer binds loosely to the incorrect target. This bond survives the temperature increase to the extension phase of the thermal cycle, triggering the creation of a copy of the intended target sequence. This copy contains the true target sequence after denaturation, and will hence be amplified in subsequent PCR cycles at the rate at which the primer amplifies its intended target. Mispriming can therefore introduce subtle effects in PCR data analysis [37]. It is sensible to investigate a calibration scheme, aimed at calibrating initial concentrations for this effect.

Let $e_{i,j}$ be the probability of producing product j from source i and c_t^{i-j} be the observed values for the threshold time until product j attains a reference quantity. Consider first two primers, one targeting 103_A and the other 103_C. After one cycle,

a sample with C at the relevant location will produce a erroneous product with A, the amount depending on $e_{C,A}$ and the initial concentration of C:

$$A(1) = e_{C,A}C(0) \tag{3.16}$$

After n cycles of amplifying this spurious product, the amount of product A is:

$$A(n) = e_{C,A}\left(1 + m + m^2 + \ldots + m^{n-1}\right)C(0) \tag{3.17}$$

$$A(n) = e_{C,A}\left(\frac{1-m^n}{1-m}\right)C(0) \tag{3.18}$$

where the last simplification uses the formula for a geometric series. Since $m < 1.75$ and $c_t > 10$, for typical values of m and c_t, $m^{c_t} - 1 > 200 - 1$, it follows that $m^{c_t} - 1 \approx m^{c_t}$, and the expression for A in (3.18) reduces to:

$$A\left(c_t^{C-A}\right) = e_{C,A}\left(\frac{m^{c_t^{C-C}}}{m-1}\right)C(0) = P_t \tag{3.19}$$

Using the same multiplication rate m, and the same initial concentration of C for a primer targeting C at the relevant location, the following also holds:

$$C(0)m^{c_t^{C-C}} = P_t \tag{3.20}$$

and therefore :

$$e_{C,A} = \frac{(m-1)(m^{c_t^{C-C}})}{m^{c_t^{C-A}}} \tag{3.21}$$

When a primer is added to samples of real DNA, arbitrary initial concentrations of sequences with A, C, G or T may exist at a relevant location. For this reason, we have to consider all primer/plasmid permutations, with one channel of faithful and three channels of spurious amplification for each primer.

$$P^i(n) = m_{i,i}^n P^i(0) + \sum_{j\neq i} e_{ji}(1 + m_{i,i} + m_{i,i}^2 + \ldots + m_{i,i}^{n-1})P^j(0) \tag{3.22}$$

Equation (3.22) can be written in a convenient form when a threshold amount of product is attained:

$$P_t = P^i(c_t^i) = m_{i,i}^{c_t^i} P^i(0) + \left(\frac{m_{i,i}^{c_t^i} - 1}{m_{i,i} - 1}\right)\sum_{j\neq i}\frac{(m_{i,i}-1)(m_{i,i}^{c_t^{j-j}})}{m_{i,i}^{c_t^{j-i}}}P^j(0) \tag{3.23}$$

$$P_t = m_i^{c_t^i}\left(P^i(0) + \sum_{j\neq i}\frac{m_{j,j}^{c_t^{j-j}}}{m_{i,i}^{c_t^{j-i}}}P^j(0)\right) \tag{3.24}$$

where,

- $i, j \in \{A, C, G, T\}$.
- The last simplification follows from the use of the formula for a geometric series, together with the formula for $e_{i,j}$.
- $m_{i,i}$, $i \in \{A, C, G, T\}$, is the faithful amplification rate when a primer of type i acts on its intended target sequence, allowing for amplification rates which depend on i

The calibration of initial concentrations can therefore be stated as a linear problem, with four equations and four unknown initial concentrations. The entries of the matrix are calculated from 12 sets of calibration runs from 12 plates, where 16 wells were prepared on each plate to measure all 16 channels of potential mispriming. Below we list the all the information required to construct the calibration matrix C, followed by a discussion. The structure is the same in all three matrices: The leftmost column is an index for the primers used $(103_A, 103_C, 103_G, 103_T)$, and the top row is an index for the plasmids.

$$m_{i,j} = \begin{bmatrix} & PlasA & PlasC & PlasG & PlasT \\ 103_A & 1.73 & 1.50 & 1.52 & 1.58 \\ 103_C & 1.62 & 1.71 & 1.47 & 1.63 \\ 103_G & 1.61 & 1.49 & 1.68 & 1.51 \\ 103_T & 1.57 & 1.63 & 1.60 & 1.71 \end{bmatrix} \tag{3.25}$$

$$c_t(i,j) = \begin{bmatrix} & PlasA & PlasC & PlasG & PlasT \\ 103_A & 20.30 & 31.05 & 27.81 & 27.57 \\ 103_C & 24.47 & 15.33 & 32.44 & 22.13 \\ 103_G & 24.79 & 31.73 & 17.06 & 34.54 \\ 103_T & 28.51 & 22.22 & 24.58 & 15.91 \end{bmatrix} \tag{3.26}$$

$$C = \begin{bmatrix} & PlasA & PlasC & PlasG & PlasT \\ 103_A & 1 & 0.04 & 0.05 & 0.01 \\ 103_C & 0.01 & 1 & 0 & 0.06 \\ 103_G & 0.4 & 0 & 1 & 0.03 \\ 103_T & 0.03 & 0.06 & 0 & 1 \end{bmatrix} \tag{3.27}$$

From the m_{i-j} matrix we note high diagonal entries, and lower rates for off diagonal entries, as can be expected for intended primer and plasmid combinations. The c_t^{i-j} matrix has low diagonal and higher off-diagonal entries, again as we expect. From the c_t^{i-j} matrix and the calculated calibration matrix C, it seems that the only channels of detectable mispriming are: $e_{T,C} = 0.03$,$e_{A,G} = 0.14$,$e_{T,G} = 0.02$,$e_{A,T} = 0.06$ and $e_{C,T} = 0.04$.

3.6 Estimating sub populations of viral genotypes

The theory must now be applied to estimate the relative concentrations of viral subpopulations $103_A, 103_C, 103_G$ and 103_T. An automated system was developed in *Matlab* to process PCR data from 253 wells on 24 plates, and to collect this data into profiles (viral load measurements over time) for the trial attendees. The functionality of the software is explained below:

- All PCR samples were processed by *NICD*, using the *ABI Prism 7000* Real Time PCR machine. Software that accompanies this machine, was used to export comma separated files, containing passive and reporter dye fluorescence. A system of *Matlab* routines was developed to process this data set.
- The first correction is a between well normalization of the reporter signal (SYBR), dividing it by the passive signal (ROX), to correct for non-PCR related fluctuations. This signal is called *Rn*.
- An exponential curve is fitted to *Rn*, and the background signal A is defined to be the average of the first ct_A cycles, where ct_A is the cycle number when the estimated curve rises to above 1 percent of the average Rn signal during the first few cycles. The signal is now called *Delta Rn*.
- Initial concentrations for 103_A,103_C,103_G and 103_T, in the arbitrary units of 'fluorescence':

$$P^i(0) \quad = \quad P_T^i e^{-\int_0^{c_t^i} \lambda_i(c)\, dc} \tag{3.28}$$

  where $i \in \{103_A, 103_C, 103_G, 103_T\}$ and λ_i is the regression line for the declining amplification efficiency, as estimated in Section 3.4.
- The initial concentrations $P^i(0)$ are normalized to 1 to obtain 'naive fractions':

$$f^i \quad = \quad \frac{P^i(0)}{\sum_j P^j(0)} \tag{3.29}$$

- These fraction are now corrected for mispriming effect. 'Calibrated fractions' (f_c^i) are obtained as the solution of the following linear problem:

$$\overline{f_c^i} \quad = \quad C\overline{f^i} \tag{3.30}$$

  C is the calibration matrix estimated in Section 3.5, $i \in \{103_A, 103_C, 103_G, 103_T\}$, and f_c and f are the naive and calibrated fractions (as vectors) respectively.

This algorithm was applied to all the PCR plates, on sets of four wells, grouped by trial attendee name and date. An example of a plot, for the processing of the samples for Patient $R8$ at 7 months, is displayed in Figure 3.9. The PCR curves for 103_A,103_C,103_G and 103_T are shown in red, green, blue and amber respectively, and the fitted amplification rates are shown in black. Vertical lines indicate when threshold cycle c_t was reached. In the title of the plot is displayed: 1) fractions calculated using a constant amplification rate of 2 (the industry standard), 2) naive fractions using equation (3.29), and 3) calibrated fractions using equation (3.30).

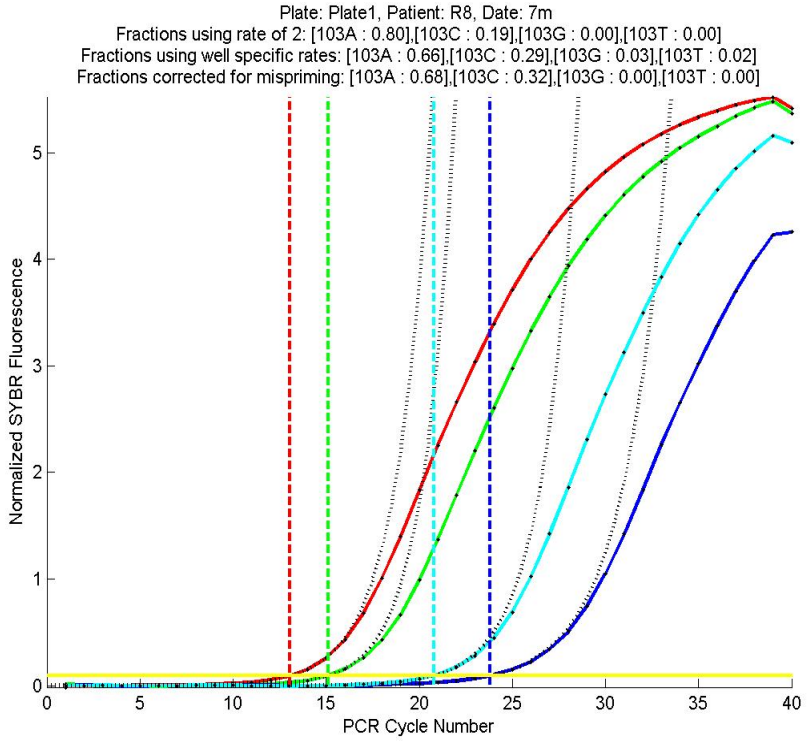

Fig. 3.9: Estimated viral subpopulations for patient 8 at 7 months

It can be noted that the industry method appears to overestimate the concentration of the strain/fluorescence that reaches threshold first. Also, that calibrating for mispriming effects removes small concentrations of subpopulations, which would be incorrectly introduced as true sub populations when 'naive' fractions are reported. Initial concentrations are converted to a resistance profile, displayed below. Figure 3.10 depicts a resistance profile for patient 8. All resistance data is pooled in Figure 3.11 . No attempt is made to indicate an 'average' profile, as it is not given by averaging over all individual profiles. The method of hierarchical Bayesian modelling is used in Chapter 5 to estimate dependence between individual parameters. The plot seems to indicate a high resistance initially, which gradually fades away as the wild type regains its dominance.

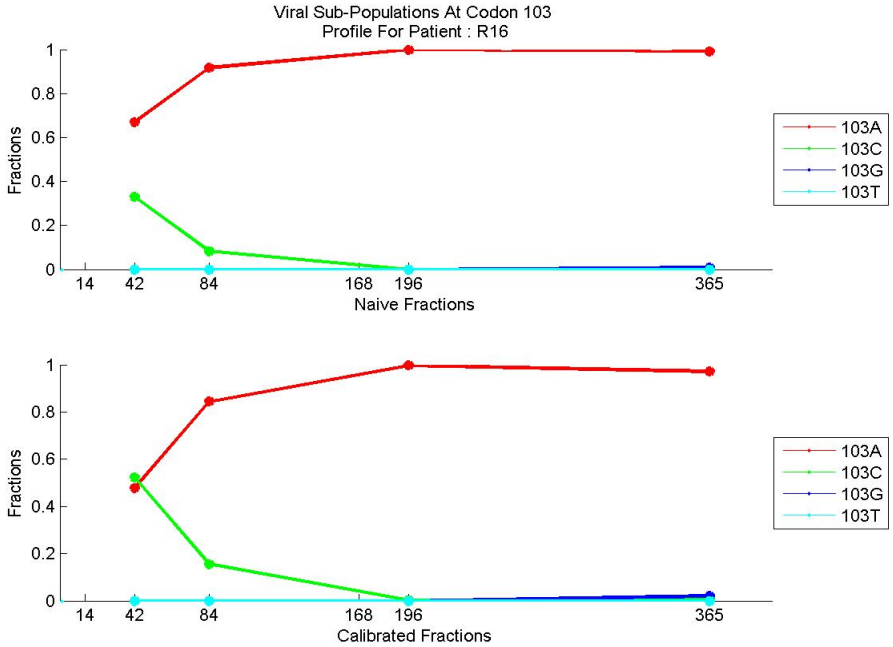

Fig. 3.10: Estimated resistance profile for patient R_{16}.

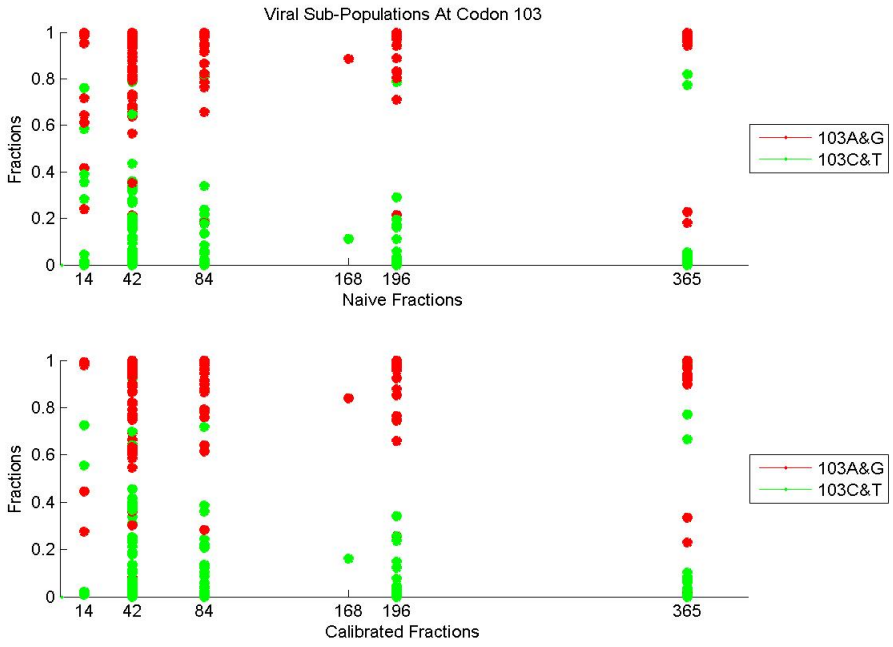

Fig. 3.11: Resistance dataset for all patients.

3.7 Repeated measurements of viral fractions

Three PCR plates were devoted in this study to repeat measurements, performing repeated PCR runs on plasma (DNA from patients) and on plasmids (commercial DNA). Figure 3.12 shows four repeated PCR runs on plasma from a patient at two weeks. It shows that significant resistant viral sub-population (103_C and 103_T) can reliably be detected at two weeks.

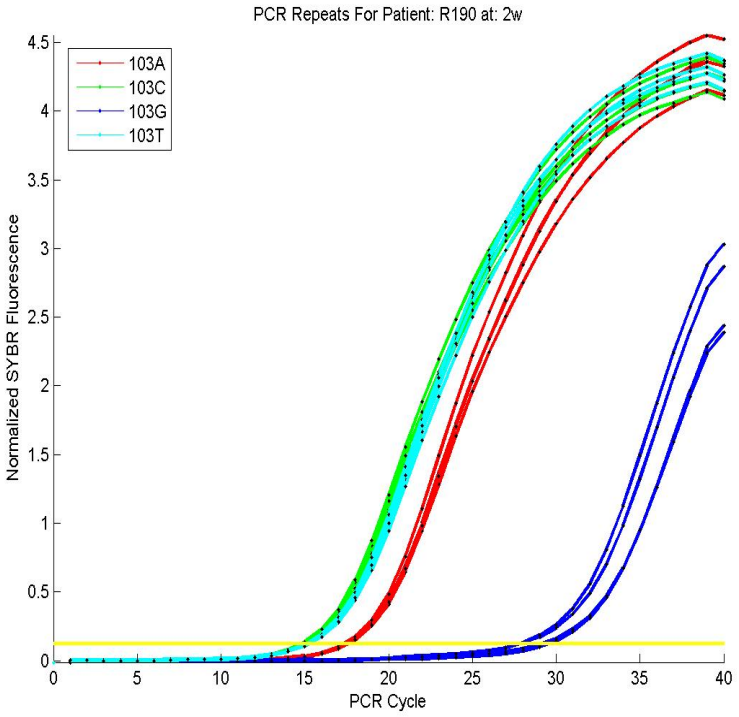

Fig. 3.12: A typical PCR repeat run. The instrument seems to reliably detect initial concentrations of different strains.

The variability of c_t is now investigated in terms of cycle number. Figure 3.13 shows the relationship of a 4-repeat c_t data with respect to the average c_t value for each repeat, and it can be seen that each c_t become more variable if threshold product is attained at an higher cycle.

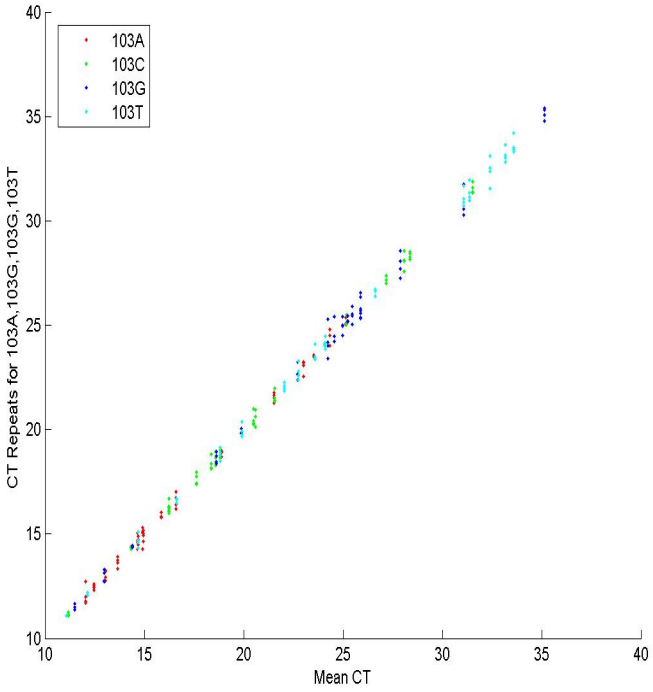

Fig. 3.13: Increase in variability in threshold detection at higher cycle numbers.

Repeatability statistics

A noise model can be derived from the set of repeated PCR runs. The algorithm outlined in the previous section can be applied to estimate the relative fraction of viral strains for each group of four measurements. The pairwise differences of all comparable (i.e each pair in a group of four measurements) relative fractions can then be pooled. A histogram of this pooling strategy is depicted in Figure 3.14.

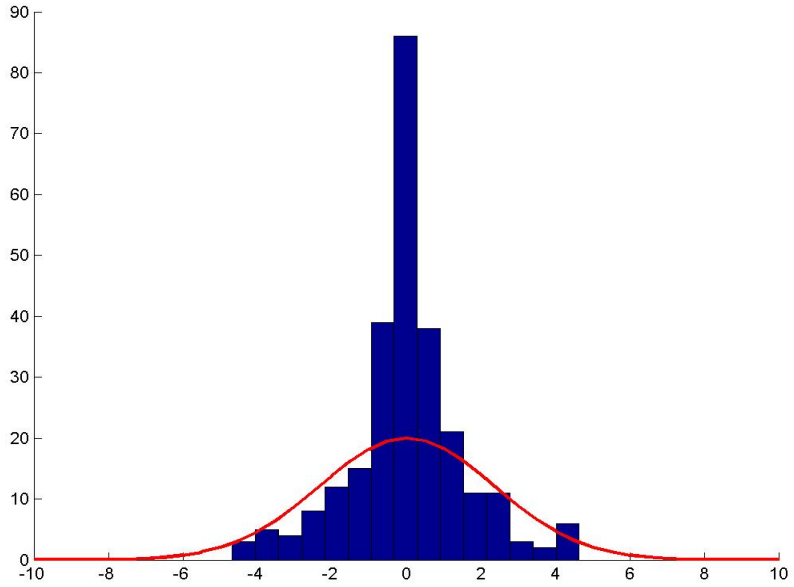

Fig. 3.14: Fitting a Gaussian noise model. This noise model does not represent the full dynamic range of the instrument, and it is at best a crude estimate of its accuracy. Many of the samples that contribute to the large bin near zero error, have been obtained 8 months after treatment. The wild type completely dominates such samples, as it has regained dominance, and the PCR can accurately detect this fact. It is unclear how well the instrument would detect a more even mix of viral strains.

It is possible to identify many factors that would contribute to the lack of repeatability of PCR processes. It is however not clear how to quantify these contributions. The factors can be broadly classified. A class of factors concerns sample preparation. Sequences of RNA are first extracted from plasma before they can be amplified. The repeats performed do not capture randomness of this extraction process. PCR experiments are usually nested, as had been the case for this particular experiment. RNA sequences must first be amplified around a codon of interest, to provide sufficient quantities for subsequent sequence-specific amplification. At each stage small but unknown pipetting errors are made. The second class contributes to the local chemistry in each well, including primer efficiencies. Another class of repeatability factors originates from the measurement device itself.

3.8 Summary

In this chapter we have seen that quantification of relative fractions for viral RNA populations by a RT-PCR process is a complex technique subject to limitations that

are inherent to PCR procedures. The method of using non-specific DNA binding dyes (SYBR in this project) is a widely used technique to quantify specific DNA product. There are known drawbacks when this method is applied to processes running at different efficiencies, as demonstrated in this chapter: 1) many well-specific and external factors contribute to different amplification efficiencies in each well, 2) at each cycle there is a contribution to these inefficiencies from primers that bind sub-optimally to their target sequences. This effect is compounded by each subsequent cycle, and 3) the primers may attach to unintended templates to produce true target sequences, an effect that is also compounded by subsequent amplification cycles.

A fundamental principle of PCR based methods is that the *average properties and behavior* of the procedure are reproducible: Each well is subjected to similar 1) concentrations of the so called master mix, 2) temperature cycles, 3) fluorescence detection, and so on. By assuming the same average process in each reaction, we can estimate viral subpopulations that are, in some average sense, free of the errors introduced by the PCR process. A model specific to each well can be used, when the PCR process is understood in greater detail. However, such a level of understanding has not been achieved in this project.

Some of the fundamental concepts of quantitative PCR analysis have been explored in this chapter, highlighting some of the factors contributing to inaccuracies. The compounding of errors associated with primer-specific detection, motivates a search for more accurate methods for using PCR to quantify relative fractions of RNA. Such methods are in fact available. The idea is to use one primer to non-specifically target and amplify a larger part of the genome. Detection of the viral subpopulations is achieved by designing probes to specifically detect codons of interest. Introducing specificity at the detection stage of each cycle leads to more accurate results, because amplification rate differences at each cycle are not compounded.

Probe-based methods are used in a wide range of applications. Its application to quantifying relative viral fractions is one of the more recent. It is unlikely to be applied extensively to data obtained from monotheraphy clinical trials, as multiple drug therapy has become the preferred method. Data from monotheraphy trials are in principle simpler to relate to models for viral evolution, as a larger modelling space can be associated with data from multi-drug therapy; i.e. there are more ways to explain the data. From a modelling perspective, a monotherapy data set would be preferred when investigating viral evolution, although this motivation is unlikely to supersede the advantages that multi-drug therapy holds for the health patient. It is therefore likely that an investigation using a probe-based method will have to deal with the fact that viral evolution will be a more hidden effect in multi-drug data.

We note the caveats and assumptions regarding this dataset, and set them aside in order to proceed with estimating parameters of the models introduced in Chapter 2. A (combinatorial) argument using the number of loaded MHC molecules on the membrane of infected cells has been used to estimate the risk an infected cell

faces of being cleared. In particular, delta-like distributions are put forward as the distributions we expect to 'see' and validate with this data set. The next chapters aim to infer more supporting evidence for this claim from the data set analyzed in this chapter.

Chapter 4

Bayesian inverse problem theory

The purpose of this chapter is to demonstrate how Bayesian inference provides a flexible and rational framework in which we can combine all the states of information associated with an inverse problem. States of information are represented by probability distributions. Information available before measurement is called the *prior* distribution. A *likelihood* distribution represents a joint state of information over the model and data space. The prior and likelihood distributions are combined using Bayes' theorem resulting in a unique *posterior* distribution. All inferences about the physical system being studied derives from the posterior distribution, and in this sense the posterior distribution represents the solution of an inference problem.

In a clinical trial designed to infer in-vivo viral parameters of different trial attendees, it is reasonable to assume that there exists a relationship between viral parameters of different individuals. Individual parameters can be viewed as independent draws from a population level distribution. In this study, hierarchical Bayesian modelling is investigated as a technique for introducing a *population level* distribution into multiparameter inference problem, and to obtain this population level distribution using the Bayesian paradigm, i.e to obtain its posterior distribution from a data set.

The Bayesian framework can be developed with simple concepts in probability calculus, and is shown to be an elegant, rational and scientific approach for studying a wide range of inverse problems associated with physical systems. This approach has many features which makes suitable and attractive for modelling inverse problems:

- The use of prior distributions is a coherent way of specifying information available independent of measurement.
- Hierarchical Bayesian modelling techniques provide a natural and systematic framework for combining uncertainties at the individual level into an inference of population level parameters. Specifying and solving hierarchical Bayesian models have only recently (in the last two decades) become feasible due to progress in Markov Chain Monte Carlo (MCMC) computational techniques [15, 38, 39], and the Bayesian approach has now become the most widely used

approach to solving hierarchical inverse problems.

- The end result of Bayesian inference is a posterior probability distribution over an inferred parameter. This is a unique feature of Bayesian inference, and the posterior distribution provides all relevant information on model parameters: mean values, error bars over parameter values, and so on. Traditional methods for solving inverse problems, such as the method of least squares, would typically return only an estimator for a parameter of interest along with ways of calculating a variance for the estimator.

Linear and non-linear squares methods are also widely used for solving inverse problems. It is known that the least-squares method provides a minimum-variance estimator for linear functions of observed data and prior model values. Least-squares methods are easy to implement (especially for linear or linearizable inverse problems) but lack robustness, as they are sensitive to outliers in the data set. This often makes 'minimum-variance' a bad criteria [13]. Prior information can be also incorporated into inverse problem by means of constraint-least-squares method, but is not straight forward to implement hierarchical prior information.

4.1 Distinguishing a model and data space

Consider an arbitrary 'system', and a set of parameters describing it. It is often not possible to observe the interesting parameters of the system directly, and they are distinguished from a set of *observable parameters*. The observable parameters, or observable events, follow a distribution that depends on *model parameters*, through a relationship called the *forward model*. Inverse theory is concerned with inferring or estimating the values of model parameters, given observed values for the observable parameters. The following steps can be distinguished when posing an inverse problem [13]:

- *Parametrization of the system*: Firstly, a *minimal* set of parameters to characterize the system has to be selected. The parameter set is separated into *model* parameters ($\mathbf{m}$) describing the system and observable parameters ($\mathbf{d}$) measuring the system. The following notation is used:

$$\mathbf{m} = m^{\alpha},\ \alpha \in I_M \tag{4.1}$$
$$\mathbf{d} = d^{i},\ i \in I_D \tag{4.2}$$

  where I_M and I_D are indices for the set of model and data parameters respectively.

- *Forward modelling*: A model that relates model parameters to data parameters is called a *forward model*. In abstract notation, it is a relation, $\mathbf{G} : \mathbf{m} \rightarrow \mathbf{d}$,

from the model space to the data space. There are many ways of finding a forward model. Relationships between model and data parameters could for instance be suggested by physical laws for a class of phenomena. It could also be a purely 'statistical' relation, obtained from accumulating correlations between the model and data space.

Phenomena typically depend on physical parameters, for example time, in complicated ways and it is therefore not always possible to describe such phenomena exactly [40]. Physical theories, or forward models, often capture only the average behavior of a system, and an important aspect of inverse theory is to describe uncertainties over inferred parameters that would be introduced by inexact forward models.

- *Inverse modelling*: The use of *observed parameters* to estimate model parameters. This is achieved by combining (states of) information on model parameters from forward modelling, information from measurements and *a priori* knowledge.

Inverse theory has wide application, such as:

- As *in vivo* viral dynamical systems cannot be observed directly, virologists usually rely on *in vitro* measurements to gain understanding about a system of virus and immune system interaction. These observation include counts of immunological markers and viral particles are usually obtained by means of *in vivo* measurements. Viral RNA, collected for example from a blood sample, cannot be counted or measured directly. The Polymerase Chain Reaction procedure amplifies small quantities of DNA *in vitro*, until it can be measured. Methods relating these measurements to estimates of in vivo viral activity were investigated in Chapter 3.

- CAT scans are used to diagnose tumors, by measuring the amount of radiation absorbed in body tissues and relating this to the opacity of the particular tissue under study.

- Geophysicists inferring properties of the earth's core from remotely sensed data.

4.2 Probability calculus

This section reviews a few fundamental concepts in probability calculus. More detailed discussions may be found in [13, 40]. Probability theory is usually developed in terms of measures over sample spaces. This rigorous formalism is not required here, and a less formal (and less concise) description of probability is used.

Let X be a non-empty set. A *measure P over X* is a function that associates a non-negative real number to any subset of A of X, in such a way that:

$$P(\emptyset) \quad = \quad 0 \tag{4.3}$$

where $\emptyset$ is the empty set and

$$P\left(\bigcup_i A_i\right) \quad = \quad \sum_i P(A_i) \tag{4.4}$$

where A_i is a sequence of disjoint sets in X. If a measure P is finite, it is called a probability distribution and is normalized to 1 in the sense that

$$P(X) = 1 \tag{4.5}$$

The Raydon-Nikodym theorem guarantees that given any probability density function P, it is possible to find a function $p(x)$, such that $P(A \subset X)$ can be expressed as an integral:

$$P(A) \quad = \quad \int_A p(x)\, d\mathbf{x} \tag{4.6}$$

where $p(x)$ is a probability density. Note that measures over sets with physical dimensions results in a density with dimension that has physical dimensions inverse to the particular volume space. A random variable is specified in terms of both the set (range) of possible values, and a measure P over this set.

The expected value of any function f(x) over X is:

$$< f(x) >_{x \in X} = \int_X f(x) p(x)\, dx \tag{4.7}$$

In this notation:

- $< X^m >$ is called the m-th moment of X.
- The mean value is $< X >$.
- The variance, or square of standard deviation, is $< (X - < X >)^2 >$.
- An important notion for multivariate distributions of x_i is the covariance matrix. The entry of the i-th row and the j-th column is given by:

$$C_{ij} = < (X_i - < X_i >)(X_j - < X_j >) > = < X_i X_j > - < X_i >< X_j > \tag{4.8}$$

Interpretation of probability

There are many ways to interpret a probability distribution, and two of the usual approaches are worth noting. The first approach is statistical. A 'random process' leads to 'realizations', also called random variables. A random variable is given a statistical description after observing many 'realizations'.

The second approach is a subjective interpretation of probability, to which the *Bayesian* approach can be applied. A particular investigator's belief of a theory or hypothesis m is treated as a random variable. When observed data d is available this belief is modified and is represented by a conditional distribution $p(m|d)$ [41]. *The most general way of describing the degree of belief or state of information on X, is to define a probability measure (and density) over X* [13]. A state of knowledge on X will generally lie between two extremes:

- *Perfect Knowledge* : When it is known that X can assume only one value, $x = x_0$, the probability density p over X is defined to be a delta function $p(x) = \delta(x - x_0)$.
- *Ignorance* is defined in terms of a 'reference' or homogeneous distribution $\mu(x)$, a controversial notion in inverse theory. A homogeneous distribution is a distribution that assigns equal probability to events occupying equal volumes in the sample space. There is no distribution, which can be called 'the homogeneous' distribution and there is no principle that can be used to define it uniquely [40]. The 'principle of insufficient reason' is often used. It implements the notion that two events must be assigned the same probability if there is no reason to believe they should be different. The hypothesis must however be accompanied by a specification for the random variables being studied. It is therefore dependent on the chosen scale, units of measurement and other properties of this specification.

There is a historic and ongoing debate about the issue of whether a probability distribution must be viewed only as a relative frequency of events, or whether it can also be viewed as a subjective state of knowledge. This discussion has a general setting [40]: can general laws be deduced from a finite number of observations?

Transformation of variables

It is often necessary to work in different parametrizations of a particular system, for example, a change of units. The Jacobian of a bijection (i.e. a one-to-one and onto map) is used to transform probability distributions.

Consider a probability measure P over a measurable space X, and a probability density p representing P. Let $y = f(x)$ be a continuous mapping of x onto another variable $y \subset Y$, representing for example a change of coordinates. The probability p_Y that y assumes a value between y and $y + dy$ is:

$$p_Y dy = \int_{y<f(x)<y+dy} p_Y \, dx \tag{4.9}$$

This can be written conveniently and equivalently as [40]:

$$p_Y dy = \int \delta[f(x) - y] p_X \, dx \tag{4.10}$$

where δ is an indicator function over the domain of $f(x)$. When f is a bijection (i.e. one-to-one and onto), the *transformation of the probability density* can be deduced:

$$p_Y(y) = p_X(x) J \tag{4.11}$$

where $J = \left| \frac{\partial x}{\partial y} \right|$ is the absolute value of the Jacobian of the transformation.

Joint, marginal and conditional distributions

Consider a two component parameter set (x_1, x_2), and its joint probability distribution $p(x_1, x_2)$. The probability that x_2 assumes a value between x_2 and $x_2 + dx_2$, while allowing x_1 to range through X_1, is called the *marginal distribution* for x_2:

$$p_{X_2}(x_2) = \int_{X_1} p(x_1, x_2) \, dx_1 \tag{4.12}$$

When the value for $x_1 = x_1^*$ is fixed, the *conditional probability* of x_2 is defined by:

$$p_{X_2|X_1}(x_2|x_1^*) = \frac{p(x_1^*, x_2)}{\int_{X_2} p(x_1^*, x_2) \, dx_2} \tag{4.13}$$

The *joint probability* is equal to the marginal probability for X_1 to have the value x_1, times the conditional probability that X_2 has the value x_2:

$$p(x_1, x_2) = p_{X_2|X_1}(x_2|x_1) p_{X_1}(x_1) \tag{4.14}$$

4.3 Inference using the Bayesian paradigm

From the symmetry in the role played by x_1 and x_2 in equation (4.13) for their joint probability distribution, follows the theorem of Bayes:

$$p_{X_2|X_1}(x_2|x_1) = \frac{p_{X_1|X_2} p_{X_2}(x_2)}{\int_{X_2} p_{X_1|X_2} p_{X_2}(x_2) \, dx_2} \tag{4.15}$$

Bayes' theorem is particularly useful in calculating a joint conditional distribution when the probability model admits a representation in terms of a parametric model and a prior distribution. It allows us to solve inverse problems: We identify the *observable parameters* $\mathbf{d} = d^i, i \in I_D$ and model parameters $\mathbf{m} = m^j, j \in I_M$ of the inverse problem. Equation (4.15) is used to infer model parameters given observable parameters:

$$p_m(m|d) \quad = \quad \frac{p_m(m)p_d(d|m)}{\int_M p_m(m)p_d(d|m)\,dm} \tag{4.16}$$

where,

- $p_m(m|d)$ is called the posterior probability distribution of the model parameters m.
- $p_m(m)$ is the prior distribution for model parameters m.
- $p_d(d|m)$, the likelihood of observing d given m, is a distribution from which the observed data can be viewed as *independent and identically distributed* (i.i.d) draws.
- The subscripts in p_m and p_d make a distinction between probability distributions used in the model and data space respectively. These subscripts will be dropped from now on, to avoid unnecessary detail in the expressions to follow. The argument of each distribution will be its domain.

Multiple parameter inference

It is often the case that an inverse problem is stated in terms of model parameters with more than one component, i.e. $m = (m_1, \ldots, m_{I_M})$, but the aim is to infer only a subset of the model parameters, say m_1. The posterior distribution for the parameter of interest is obtained by averaging the joint posterior distribution over the remaining parameters, often called *nuisance parameters*:

$$p(m_1|d) = \int p(m_1, \ldots, m_{I_M}|d)\,dm_2 \ldots d_{I_M} \tag{4.17}$$

It is often not possible to use Bayes' theorem to infer only a parameter of interest, m_1 say, directly and without reference to nuisance parameters. This occurs when the specification of the likelihood $p(d|m_1, m)$ includes the nuisance parameters m. Such a likelihood would typically be interpreted as a distribution from which observed data are i.i.d given m_1 *and* the nuisance parameters.

If the likelihood does not depend on a parameter of interest m_1, then m_1 is said to be *non-identifiable* from the data [41]. Suppose $m = (m_1, m_2)$ and that likelihood is given by $p(d|m) = p(d|m_2)$, i.e d is independent of m_1 given m_2. From

the marginal distribution for m_1 it can be seen that d provides information on m_1 indirectly through m_2 if we assume knowledge of $p(m_2|m_1)$:

$$p(m_1|d) \propto p(m_1) \int p(m_2|m_1)p(d|m_2)\, dm_2 \tag{4.18}$$

Hierarchical modelling

An important class of inverse problems involves model parameters that are related in some way. For example, in a clinical trial designed to infer in-vivo viral parameters of different trial attendees, it is reasonable to assume that there exists a relationship between viral parameters of different individuals. It is natural to view these individual parameters as draws from the same population level distribution. Clinical trials are in fact designed to inform strategies for the population represented by the trial attendees. Estimating the population level distribution is the primary motivation for the trial. The use of a population level distribution is called hierarchical Bayesian modelling. A population distribution is also called a *hyperdistribution*, and is described in terms of *hyperparameters*. This section investigates the hierarchical modelling strategy.

Suppose that nothing is known about the hyper distribution. After obtaining posterior distributions for a few individuals, one will be able to provide constraints for the hyper distribution. And generally, as more individual posterior distributions become available, the description of the hyper distribution will become more precise. Does it make sense, from a Bayesian inference point of view, to use this hyper distribution as a prior for inference regarding subsequent observed individuals? Objections to this approach include:

- The same data set would be used twice, a procedure that cannot increase the precision of posterior distributions.
- Would the prior distribution change if we changed the order in which posterior distributions are combined?

The approach advocated by [42, 38] is to estimate the population or hyper distribution from all data. Assume that a set of parameters $m_1, \ldots, m_{I_M}$ represents viral parameters for individual patients, and also that the population level distribution is given by $p(\phi)$. The distribution p could for example be a normal distribution with a known variance and an unknown mean ϕ. The following figure, adapted from [38], is a useful way of visualizing individual parameters as draws from a population level distribution:

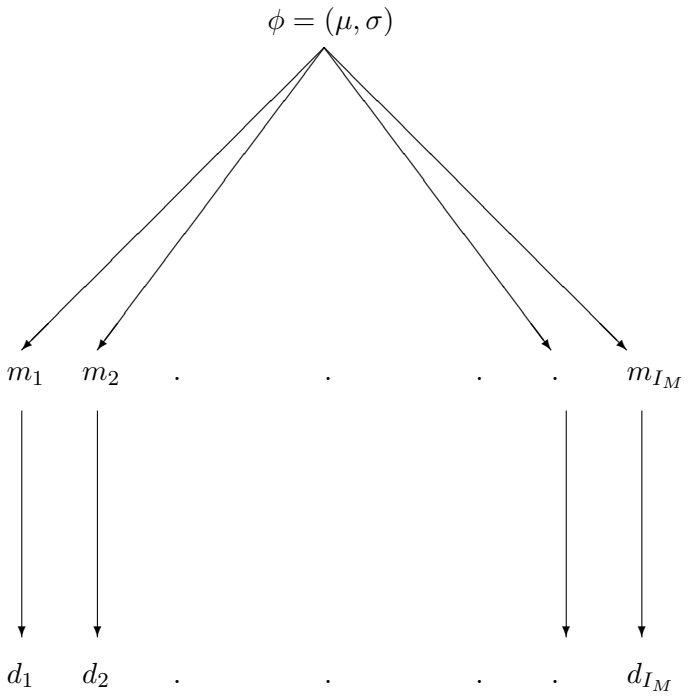

Fig. 4.1: A simple hierarchical model. The hyperparameter $\phi = (\mu, \sigma)$ typically represents the mean and variance of the population level distribution. The point emphasized by this model is that data d_i is obtained for each individual and this information must be propagated to the population level distribution $\phi = (\mu, \sigma)$ via individual parameters m_i.

Individual parameters can be assumed to be exchangeable when no information is available to suggest otherwise. This means that they are independent and identically distributed given ϕ. Their joint probability distribution has the form:

$$p(m_1, \ldots, m_{I_M}|\phi) \quad = \quad \prod_{i=1}^{I_M} p(m_i|\phi) \tag{4.19}$$

and is obtained by averaging over ϕ, which may be treated as nuisance parameters:

$$p(m) \quad = \quad \int \left[\prod_{i=1}^{I_M} p(m_i|\phi) \right] p(\phi)\, d\phi \tag{4.20}$$

The set of random variables $(m_1, \ldots, m_{I_M})$ is called exchangeable if the joint distribution $p(m_1, \ldots, m_{I_M})$ is invariant with respect to permutation of the labels

$(1, \ldots, I_M)$. A theoretical result, known as *de Fenetti's theorem* states that if the sequence of real-valued random variables is exchangeable, then there exists a parameter θ and a distribution $p(\theta)$ that allows a rewriting of the joint distribution in the form of equation (4.20). The existence of $p(\theta)$ according to this theoretical result is often used for a justification of both its interpretation and its use as a prior distribution. This interpretation is debated [43], where objections to this interpretation include the fact that the theorem requires an infinite set of exchangeable variables.

A hierarchical model involving ϕ, m and d can be given the following general three stage specification [38, 42, 41]:

$$p(\phi, m, d) = p(\phi)p(m|\phi)p(d|m) \tag{4.21}$$

which requires definitions for:

$$p(d_1, \ldots, d_{I_M}|m_1, \ldots, m_{I_M}) = \prod_{i=1}^{I_M} p(d_i|m_i) \tag{4.22}$$

$$p(m_1, \ldots, m_{I_M}|\phi) = \prod_{i=1}^{I_M} p(m_i|\phi) \tag{4.23}$$

$$p(\phi) \tag{4.24}$$

where,

- At the first stage of the hierarchy one specifies the parametric form of the likelihood model. The data set $d = d_1, \ldots, d_{I_M}$ represents observables for I_M related sources.
- At the second stage, individual parameters $m_1, \ldots, m_{I_M}$ are modelled as exchangeable, given the population-level or hyperparameter ϕ. A hyperparameter often directly characterizes the population level distribution. It may, for example, represent the mean or the variance of the population level distribution from which individual parameters $m_1, \ldots, m_{I_M}$ are drawn.
- Finally, a prior distribution $p(\phi)$ is specified for the hyperparameter. The hyperprior distribution defines a prior for m via exchangeability relationship (4.20).

Bayes's theorem is used to make inferences about $m_1, \ldots, m_{I_M}$ and ϕ:

$$p(m_i|d) = \int p(m_i|\phi, d)p(\phi|d)\, d\phi \tag{4.25}$$

where

$$p(m_i|\phi, d) \propto p(d|m_i)p(m_i|\phi) \tag{4.26}$$
$$p(\phi|d) \propto p(\phi)p(d|\phi) \tag{4.27}$$
$$p(d|\phi) = \int p(d|m)p(m|\phi)\, dm \tag{4.28}$$

The last two equations show that the hyperparameter ϕ is non-identifiable with respect to the data d in the sense of (4.18). Evaluation of equation (4.28) appears to be a daunting task, as it is multiple integral with the following form:

$$p(d|\phi) = \int_{L(m_1)}^{U(m_1)} \dots \int_{L(m_{I_M})}^{U(m_{I_M})} \prod_{i=1}^{I_M} p(d_i|m_i) \prod_{i=1}^{I_M} p(m_i|\phi)\, dm_1, \dots, \, dm_{I_M} \tag{4.29}$$

where $L(m_i)$ and $U(m_i)$ are the lower an upper bounds in the range of integration for parameter m_i of the joint distribution $p(m) = p(m_1, \dots, p_{I_M})$. The integrand can be written as product of I_M terms that depend on each of the variables $m_1, \dots, m_{I_M}$ respectively. This multiple integral can be written as a product of integrals using Fubini's theorem:

$$\begin{aligned} p(d|\phi) = & \int_{L(m_1)}^{U(m_1)} p(d_1|m_1)p(m_1|\phi)dm_1 * \dots * \\ & \int_{L(m_{I_M})}^{U(m_{I_M})} p(d_{I_M}|m_{I_M})p(m_{I_M}|\phi)\, dm_{I_M} \end{aligned} \tag{4.30}$$

4.4 Solving inverse problems by combining states of information

In Section 4.3 we have demonstrated that the unique solution of an inverse problem is given by a posterior distribution over model parameters. It is the result of combining all available information, namely information represented by the definition of a prior distribution and information obtained from measurement. The purpose of this section is to investigate the notion of information in inverse problems. More details regarding information in inverse problems and their classification presented in the next few sections, may be found in [13, 39].

Relative information content

The following measure can be used to make a precise statement about relative information content. Given two normalized density functions, the *relative information content* of $p_1(x)$ with respect to $p_2(x)$, is defined by:

$$I(p_1, p_2) = \int p_1(x) \log \left(\frac{p_1(x)}{p_2(x)} \right) \, dx \tag{4.31}$$

This is also known as *Shannon's measure* or *entropy* of information content. The relative information content of $p(x)$, with respect to a homogeneous probability density $\mu(x)$, is called the information content of $p(x)$. Information content has the properties:

- The information content of a homogeneous distribution with respect to itself is zero. All other distributions have positive information content with respect to a homogeneous distribution.
- The relative information content of a distribution is invariant under a bijection. In this sense, two proportional distributions are equivalent.
- A sharply peaked distribution has high information content relative to flat or diffuse distributions.

These properties make this measure suitable for measuring how much information a prior distribution and a likelihood function introduce into an inverse problem respectively *when compared to a homogeneous probability density*. The difference between the information content of the prior and posterior densities is a measure of the amount of information introduced by experiment.

Shannon's information measure is also used to derive prior information that would maximize entropy under certain constraints. This is the basic idea of the principle of *maximum entropy*, discussed briefly in Section 4.4. This principle 'selects' a distribution that would contain *minimum information*, apart from the information introduced by specific constraints.

Information in inverse problems

A good understanding of all the uncertainties (states of information) associated with a particular problem is necessary for a successful application of any inverse theory. In this section a few idealized examples are used to explain such information.

Information from forward modelling

Generally, forward models cannot predict data exactly, due to modelling and experimental errors. These two sources of uncertainties are described by a joint probability distribution, $\theta(d, m)$, on the data and model space. It gives the probability that the forward model will simultaneously predict the model and data parameters. This distribution can be factored as

$$\theta(d, m) \quad = \quad \theta(d|m)\mu_M(m) \tag{4.32}$$

when an explicit forward model is available.

- For an exact theory:

$$\theta(d|m) \quad = \quad \delta\left(d - g(m)\right) \tag{4.33}$$

- It is often necessary to treat uncertainty introduced by inexact forward models. This uncertainty is defined by a probability distribution. A number of reasons for an inexact theory may be distinguished [39]:

  – The model predicted data values may include a statistical element. These models g are stochastic forward models, such as models given by stochastic differential equations. Other examples include models that are approximations to exact models.

  – Uncertainty in model predicted values may be introduced due to uncertainty in specifying model parameter values m in the forward model $g(m)$.

  – Uncertainty in g could result from uncertainty in specifying $g(m)$ and m.

  The following form is often used for an inexact theory that assigns independent 'error bars' to predicted data values:

$$\theta(d|m) = \left[\Pi_{i\in I_D}\frac{1}{2\sigma^i(m)}\right]\exp\left[-\Sigma_{i\in I_D}\frac{|d^i - g^i(m)|}{\sigma^i(m)}\right] \tag{4.34}$$

  where I_i is an index for the data set d and $\sigma^i(m)$ is the precision of the i-th data point that depends on the value of the model parameter m.

  When error bars follow a Gaussian distribution, uncertainties in the theory assume the following form:

$$\theta(d|m) = \frac{1}{(2\pi)^N det(C_T)}\exp\left[-\frac{1}{2}(d-g(m))^T C_T^{-1}(d-g(m))\right] \tag{4.35}$$

  where C_T is the covariance operator over the model space.

- When errors are *independent of the model parameter values*:

$$d = g(m) + \epsilon_T \tag{4.36}$$

  and

$$\theta(d|m) = f_T(d - g(m)) \tag{4.37}$$

  where f_T describes the error statistics for the theory.

Information obtained from measurements

Instruments cannot measure observable parameters exactly and repeated measurements will vary in a random way, even under strict experimental control. This state of information is described by the conditional probability density, $v(d_{obs}|d)$, of observing d_{obs} given that the true value is d. The forward modelling errors $\theta(d|m)$ reported in the previous section can alternatively be stated as measurement errors $v(d_{obs}|d)$.

It is in also possible to treat forward modelling and measurements errors separately. A frequently used assumption in inverse problems is that measurement errors are independent from measurement input. Also, that errors in the theory $\theta(d|m)$ are independent of the model values m. These assumptions can be stated as:

$$d_{obs} = d + \epsilon_D \tag{4.38}$$

and

$$d = g(m) + \epsilon_T \tag{4.39}$$

respectively. It is also assumed that these errors follow known distributions:

$$v(d_{obs}|d) = f_D(d_{obs} - d) \tag{4.40}$$

and

$$\theta(d|m) = f_T(d - g(m)) \tag{4.41}$$

If it can be assumed that measurement and theoretical uncertainties are uncorrelated, then applying Bayes' theorem gives [39, 13]:

$$p(m|d) = \frac{f(d_{obs} - g(m))p_M(m)}{\int f(d_{obs} - g(m))p_M(m)\,dm} \tag{4.42}$$

where

- $\epsilon = \epsilon_D + \epsilon_T$ and $f(\epsilon) = f_D(\epsilon) * f_T(\epsilon)$ is the convolution of f_D and f_T.
- $p_M(m)$ is a prior distribution over model space M.

This demonstrates how the likelihood function can be constructed when uncertainties in measurement and theory can be distinguished. The convolution of theoretical and experimental uncertainties given by equation (4.42) is particularly simple to calculate when both follow Gaussian probability distributions. In this case the result is again Gaussian.

A Priori information

The prior distribution is a distribution $\rho_M(m)$ over the model space that is independent of measurements d. Information on a parameter is often available before measurement. The order of magnitude and bounds for m are usually known. It could be the result of previous measurement, or it may be implied by the physical laws of the class of phenomena being studied. Inserting specific information into a prior distribution, results in informative prior distributions. Maximum entropy methods are used to obtain such priors. This section reviews some of the frequently used prior distributions.

Flat or uniform priors

If constraints for a quantity m can be given only in terms of upper and lower bound for model parameter values, a *uniform distribution* is used:

$$p_M(m) = \begin{cases} \frac{1}{(U_m - L_m)} & : \quad L_m \leq m \leq U_m \\ 0 & : \quad \text{otherwise} \end{cases} \tag{4.43}$$

where L_m and U_m are the lower and upper bounds of m respectively.

Log-normal priors

The log-normal distribution is often used to model a priori information based on the constraint that a physical parameter must be positive. In this case, the log of the parameter follows a normal (Gaussian) distribution.

Minimally informative priors

Many approaches have been developed to find prior distributions that would affect the posterior distribution in a minimal way. A minimally informative distribution is also called noninformative or vague. The principle of *transformation invariance*, introduced by Jeffreys, leads to such priors. According to this principle, a rule or requirement used to select a prior distribution for a parameter, must give the same result when the parameter is subjected to a one-to-one transformation. If a prior density $p(m)$ for m is selected, then a prior density for a transformed parameter, $m^* = h(m)$, must be consistent with equation (4.11):

$$p(m^*) \quad = \quad p(m) \left| \frac{\partial m}{\partial m^*} \right| \tag{4.44}$$

This principle leads to the use of [44, 38]:

- $p(m) \propto K$ for a *location parameter*, e.g. the mean of a normal distribution with known variance.

- $p(m) = \frac{K}{m}$ for a *scale parameter*, e.g. the variance of a normal distribution with a known mean. The value of K is determined by a normalization requirement for m.

Another approach to obtain noninformative priors, also suggested by Jeffreys, is based on Fisher's information content. This principle chooses

$$p(m) \quad = \quad [J(m)]^{\frac{1}{2}} \tag{4.45}$$

as a noninformative prior, where

$$J(m) = - \left\langle \frac{\partial^2 log\, p(x|m)}{dm^2} \middle| m \right\rangle \tag{4.46}$$

is the Fisher information for m. It follows easily that $J(m)$ is transformation invariant [38]. Using equation (4.45), it can be demonstrated [44] that the prior for a location θ and a scale parameter σ is given by:

$$p(\theta, \sigma) = \frac{1}{\sigma^2} \tag{4.47}$$

Priors from maximum entropy arguments

The principle of *maximum entropy* can be used to introduce *specific information* into a prior distribution so that it would reflect the current state of knowledge. For example, if the moments M_k of a probability density function $p(x)$ are known, then the functional form of $p(x)$ can be obtained by minimizing an objective function: Find the distribution $p(x)$, with minimal information content and the given moments M_k. It is solved using Lagrange multipliers (λ_k):

$$\int p(x) \log \frac{p(x)}{\mu(x)}\, dx + \Sigma_k \lambda_k \left[\int x^k p(x)\, dx - M_k \right] \tag{4.48}$$

The following is an brief account of familiar maximum entropy distributions [39]. Each case is subjected to additional constraints on the moments M_k of p, where

$$M_k = \int x^k p(x)\, dx \tag{4.49}$$

- The principle of maximum entropy selects the uniform distribution when the only requirement is that a distribution should be normalized in an interval $[a, b]$.
- A second case arises when, in addition to the normalization requirement, the expected value of p in $[a, b]$ is known, i.e $M_1 = m$. It can be shown that the principle of maximum entropy selects a distribution with an exponential form.

- When the first three moments, i.e M_0,M_1 and M_2 of p are known, maximum entropy selects a Gaussian distribution, making this distribution admissible as a prior in a variety of inference problems.

Hierarchical priors

A higher order distribution $p(m^*)$ can be introduced to model the dependence among the model parameters, when it is known that model parameters are related in some way. Higher order distributions are usually called hyper or hierarchical priors and they are used in hierarchical probability models, as discussed in Section 4.3. For these models a joint prior distribution is introduced:

$$p(m, m^*) \quad = \quad p(m^*)p(m|m^*) \tag{4.50}$$

Hierarchical modelling is used in Section 5.3, where viral parameters of a particular trial attendee are viewed as draws from a 'population level' distribution. The main purpose of a clinical trial is indeed to understand this hyper distribution, and to design treatment strategies for the 'population' it represents.

For some problems, a higher order distribution can often be derived by intuitive arguments. Consider an experiment designed to estimate the half-life of an arbitrary element. A natural prior distribution to use is the distribution representing the half-life of all known atoms. The half-life of the particular atom must clearly be a draw from this distribution. In an interesting example, [13] derives this distribution by constructing a histogram from the tabulated half-lives of (most) known atomic nuclei, showing it to be in good agreement with log-uniform distribution derived using invariance arguments. We may similarly expect the hyper-parameters relevant to the estimation of in-vivo viral parameters to be tabulated in future as more results from these types of inverse problems become available.

Conjugate priors

The requirement that the posterior distribution follow the same distribution as the prior distribution is called *conjugacy*. This method is often used to simplify calculations when the analytic form of the prior distribution and the likelihood model are known. Although it might seem questionable to choose conjugate priors just for the sake of 'simplifying' calculations, their use can usually be justified when using a dominant data set, i.e a data set that would give a likelihood function $p(d|m)$ which dominates the prior distribution $p(d|m)$ when applying Bayes' theorem to calculate a posterior distribution.

Choosing a prior distribution

What is a natural prior to use? It depends on what is meant by natural, and what information is available:

- The uniform or log-normal distribution can be used when physical constraints of the system are known.
- The method of maximum entropy is used when specific information has to be introduced into the prior distribution.
- Conjugate priors are used when it is important obtain a simple analytical form for the posterior distribution.
- Hierarchical priors are used when it is natural to view parameters as draws from a population level distribution. These (hyper) priors are also preferred if the probability model contains a large number of parameters, and they are often chosen to meet conjugacy requirements.
- Jeffrey's invariance principles can be used to find priors that are minimally informative. This method is usually not used when a population level description for the prior distribution is available.

There are clearly no cast-in-stone rules for selecting the priors for a particular inverse problem. What has to be borne in mind, even when using the best possible prior distribution, is that solving an inverse problem is a passage from prior to posterior information, obtained by conditioning on observed data. An informative data set will dominate any reasonable prior distribution. The choice of prior should ultimately be a statement about what kind of prior information for a particular inference problem would make the data dominant [42].

4.5 Summary

Some of the fundamental concepts in Bayesian inference theory have been discussed in this chapter. Following [13], an inverse problem has been divided into three steps: 1) distinguishing a model and data space 2) obtaining a forward model from the model to the data space and 3) estimating model parameters from observed values of observable parameters.

States of information over model and observable parameters were defined by probability distributions, and an inverse problem was described as combining a priori information, theoretical information and information obtained from measurement. The notion of combining information to solve inverse problems is formalized by [13] and is reported in Appendix G.

Hierarchical Bayesian modelling techniques are used to construct probability models that are required to accommodate individual parameters with population level correlations. This technique is widely used to interpret longitudinal data sets, where models must allow within-subject and between-subject variations. Instead of using ad hoc methods to find a reasonable prior for individual parameters, it was argued that a population distribution must be used, from which individual parameter can be viewed as 'iid' draws. This population level distribution defines the prior

distribution for individual parameters, and its posterior distribution is obtained by performing Bayesian inference on the joint distribution of model parameters and on the complete data set. Hierarchical modelling poses computational challenges on all but the simplest models. Advanced computational techniques, such as Markov Chain Monte Carlo (MCMC) methods [15, 38, 39] are used when high levels of processing are required.

Chapter 5

Estimating viral parameters

The purpose of this chapter is to develop and implement procedures for inferring viral parameters by fitting multi-strain resistance models developed in Chapter 2 to the longitudinal data set for resistant viral fractions analyzed in Chapter 3.

The objective is to infer parameters of individual cells, such as lifetime distributions, based on observing populations of infected cells. To this end, models for the competition and turnover dynamics of two strains have been implemented in Section 2.2. These models incorporate survival probabilities of *individual* cells as 'the number of stages' that infected cells pass through, into a model for *populations* of infected cells given by the equations (2.46) to (2.48). Thus, information about individuals can be obtained from measurements on a population of such individuals, by structuring properties of the individual into equations for the population. These parameters are inferred using the Bayesian framework developed in Chapter 4. Inferring 'the number of stages' will shed some light on the shape of the lifetime distributions of infected cells. Inferring other parameters, where the focus will be on treatment efficiency and duration, will show that a class of widely used models will fit clinical data only when these parameters assume unrealistic values.

These objectives are addressed by solving an inverse problem.

5.1 Statement of the inverse problem

An inverse problem can be formulated as the combination of a priori, theoretical and experimental information, stated in [13] and explained in Section 4.3:

$$p(m|d) = \frac{p(m)p(d|m)}{\int_M p(m)p(d|m)dm} \tag{5.1}$$

where,

- $p(d|m)$ represents theoretical and measurement information. The host-viral dynamical systems are modelled by deterministic models and they are assumed to be exact, i.e $p(d|m) = \delta(d - g(m))$. As a consequence of this assumption, the likelihood function can only be defined in terms of measurement errors:

$$d_{i,j} = g(m_i, t_{ij}) + \epsilon_{i,j} \tag{5.2}$$

where $d_{i,j}$ represents the measured response and $g(m_i, t_{ij})$ the predicted response for the i-th patient at time j. The usual assumption is that errors $\epsilon_{i,j}$ are normally distributed:

$$p\epsilon_{i,j}|m_i) \sim N(\mu, \tau^{-1}) \tag{5.3}$$

where τ^{-1} is the precision of measurement and μ is often taken to be zero. Approximate values for τ^{-1} are derived in Section 3.7.

- Data parameters have the form:

$$d_{i,j} = (R_i, t_{i_j}), t_{i,j} \in \{14, 42, 84, 168, 196, 365\,\text{days}\} \tag{5.4}$$

where $R_{i,j}$ represents the fraction or ratio of mRNA of patient i that originates from the resistant strain. These fractions were measured at 2 weeks, 6 weeks, 3 months, 7 months, 8 months and 12 months respectively. Complete profiles are not available for all patients, as some patients failed to complete the trial.

- The complete set of model parameters for each patient is:

$$m = \{S, \mu_T, \mu_{P_{i,j}}, \mu_V, f, k_i, N_{i,j}, n, TreatmentDuration\} \tag{5.5}$$

Innate infectivity of the wild-type and resistant viral strains k_i are indexed by i, and it is reduced, for a period of $TreatmentDuration$, by different amounts for the two competing strains. The number of stages the infected cells pass through are represented by n.

We are going to solve different inverse problems by assuming known values for all parameters, except for $TreatmentDuration$, $k_{i,j}$ and n in each of these inference model. Values for fixed model parameters are given by Table 2.1.

- $p(m)$ represents prior information on the joint data and model parameter space.

When inferring $TreatmentDuration$ and $k_{i,j}$ we will assume flat priors (explained in Section 4.4). For the parameter representing 'the number of stages' it is necessary to introduce a homogenous prior. Section 5.4 investigates two intuitive arguments that could be used to obtain $\mu(n)$.

- $p(m|d)$ represents the posterior distribution over the model parameters and the solution of the inverse problem.

5.2 Estimating treatment efficiency and duration

The purpose of this section is to see what would happen if we fit the simplest of mathematical models to the strain dynamics data set of Chapter 3. These models assume exponentially distributed lifetimes for populations of infected cells and they are a limiting case of gamma lifetime distributions introduced in Section 2.4, where only one stage in the life cycle of infected cells is assumed. Estimating treatment duration and efficiency has the following Bayesian solution:

$$p_M(R, Dur|d) = \frac{p_M(R, Dur)p_D(d|R, Dur)}{p_D(d)} \tag{5.6}$$

where,

- $p_M(R, Dur) = p_M(R)p_M(Dur)$ is the joint prior distribution for treatment efficiency or resistance R and treatment duration Dur. The prior for R is assumed to independent of the prior for Dur. Flat priors are assumed:

$$p_M(R, Dur) = \begin{cases} \frac{1}{(U_R - L_R)(U_{Dur} - L_{Dur})} & : \quad L_R \leq R \leq U_R \\ & : \quad L_{Dur} \leq Dur \leq U_{Dur} \\ 0 & : \quad \text{otherwise} \end{cases} \tag{5.7}$$

  It is assumed that lower and upper bounds for R and Dur is given by $L_R = 0, U_R = 0.6$ and $L_{Dur} = 0, U_{Dur} = 30$ respectively.

- The likelihood function $p(d|R, Dur)$ is assumed to follow a Gaussian distribution:

$$p_D(d|R, Dur) = \prod_{i=1}^{I_D} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(d_i - g(R,Dur))^2}{2\sigma^2}} \tag{5.8}$$

- The marginal distribution $p(d)$, i.e. the normalization factor for the posterior distribution in equation (5.6), is given by:

$$p_D(d) = \int_{L_R=0}^{U_R=0.6} \int_{L_{Dur}=0}^{U_{Dur}=30} p_M(R, Dur)p_D(d|R, Dur)\, d(R)\, d(Dur) \tag{5.9}$$

The solution of the inverse problem stated by equation (5.6) is given by the posterior distribution $p(R, dur|d)$ and is displayed in Figure 5.1.

It can be seen that approximately three weeks of very efficient treatment is required to produce resistance to Nevirapine, when the treatment induced selective pressure is 'perfect', i.e. if reproduction is completely blocked for the wild-type,

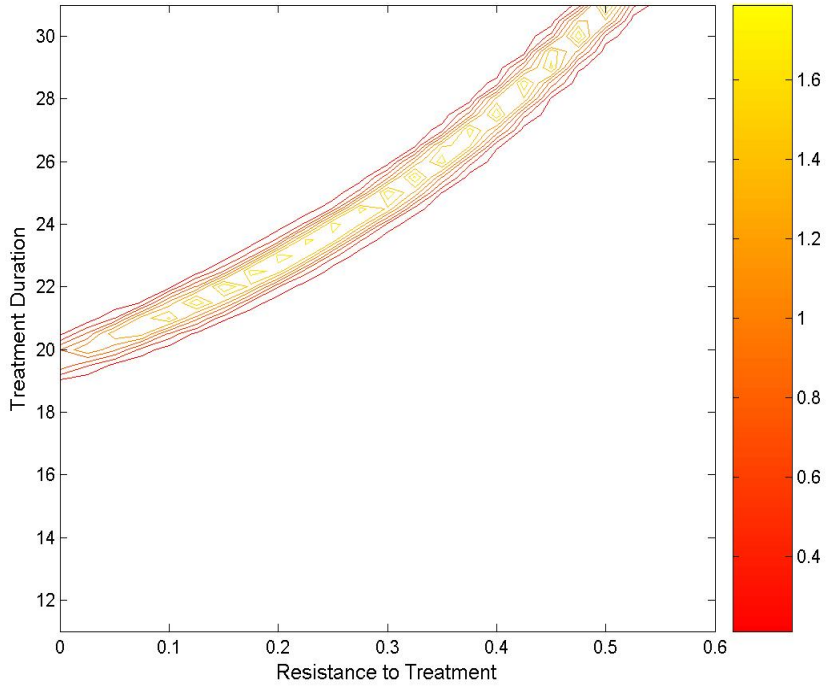

Fig. 5.1: Contour plot of joint posterior probability density for treatment efficiency and treatment duration.

but not affected at all for the resistant strain. Imperfect treatment requires a longer treatment period to produce a significant drug resistant population. Note that these model has used the same value for k_1 and k_2 for the infectivity parameter for the two competing strains. The reality, suggested by in-vitro experiments conducted to track the relative fitness of competing clones [7], is that there are significant fitness differences between competing strains. The model gives an almost unrealistic fitness advantage to the resistant strain during treatment.

The response times of the simplest models (i.e exponential lifetime distribution for populations of infected cells) are apparently to slow to exhibit realistically the fast rise in a drug resistant viral population.

5.3 Estimating the population-level distribution for treatment efficiency and duration

The conclusion of the previous section is thus far based on the viral profile for a single patient, and the question arises if this conclusion will also hold for the population this patient is drawn from. This population refers in particular to a population of Nevirapine naive women. These women are infected with the wild type $K103$ that

frequently mutates into resistant type $K103N$. A hierarchical probability model is specified as follows:

$$p(\phi|d) = \frac{p(\phi)p(d|\phi)}{p(d)} \tag{5.10}$$

where,

- $p(\phi)$ is the prior for the population level distribution parameterized by ϕ. In the implementation details introduced shortly, $\phi = (\mu_R, \sigma_R)$, i.e. we assume a normal distribution for the population level treatment efficiency R, with mean μ_R and variance σ_R.
- The likelihood function $pd|m)$ has the following general form:

$$p(d|\phi) = \int p(d|m)p(m|\phi)\, dm \tag{5.11}$$

  The hyper-parameter ϕ affects the data only through m.
- $p(d)$ is the normalization factor for the posterior distribution $p(m|d)$.

Due to complexity and time constraints in numerically evaluating high dimensional integrals, the population level distribution for R(resp. Dur) will be estimated separately, assuming a fixed value for Dur (resp. R). Advanced computational techniques, such as MCMC (Markov Chain Monte Carlo) Bayesian analysis [45, 38], are required for multidimensional hyper-distribution problems, and the details of these methods are not pursued here.

Inferring population level treatment efficiency

It is assumed that the population level distribution follows a normal distribution, i.e $p(\phi) \sim N(\mu_R, \sigma_R)$, and the objective is to estimate $\phi = (\mu_R, \sigma_R)$. The prior distribution for ϕ is specified as:

$$p(\phi) = p(\mu_R)p(\sigma_R) \tag{5.12}$$

assuming that prior information in μ_R is independent of prior information in σ_R. Referring to Section 4.4 it is seen that sensible priors for the location parameter μ_P is given by:

$$p(\mu_R) = \begin{cases} \frac{1}{(U_{\mu_R} - L_{\mu_R})} & : \quad L_{\mu_R} \leq \mu_R \leq U_{\mu_R} \\ 0 & : \quad \text{otherwise} \end{cases} \tag{5.13}$$

and a sensible prior for a scale parameter σ_R is given by:

$$p(\sigma_R) = \begin{cases} \frac{1}{\sigma_R^2} & : \quad L_{\sigma_R} \leq \sigma_R \leq U_{\sigma_R} \\ 0 & : \quad \text{otherwise} \end{cases}$$

The likelihood of observing data set d given the population level distribution $\phi = (\mu_R, \sigma_R)$, has the form given by equation (4.30). It is a a product of likelihoods of observing each individual profile:

$$p(d|\mu_R, \sigma_R) \quad = \quad L(d_1|R_1) \ldots L(d_{I_M}|R_{I_M}) \tag{5.14}$$

The population level distribution $\phi = (\mu_R, \sigma_R)$ influences the likelihood indirectly through ϕ:

$$L(d_i|R_i) \quad = \quad \int_{L(R_i)}^{U(R_i)} p(d_i|R_i) p(R_i|\mu_R, \sigma_R)\, dR_1 \tag{5.15}$$

where $i \in (1, \ldots, I_M)$ and d_i and R_i represent the observed resistance profile and the treatment efficiency parameter R, experienced by patient i. The likelihood of observing each profile $p(d_i|R_i)$ is given by equation (5.8). The posterior distribution $p(\mu_R, \sigma_R|d)$ is obtained by substituting numerical evaluations of expressions (5.12) - (5.15) into equation (5.10).

Due to the lack of actual data to implement and test a hierarchical inversion algorithm, a hypothetical data set from a hypothetical trial was generated. To this end, a normal distribution for the population level distribution for treatment efficiency is assumed. The distribution was given mean of 0.2 and a variance of 0.1 and 100 samples were drawn from this distribution. Viral profiles were generated based on these values for treatment efficiency. Gaussian noise, with a mean of zero and a variance of 5 percent, was added to these profiles to produce the data set displayed in Figure 5.2. The resulting posterior distribution is displayed in Figure 5.3 and Figure 5.4.

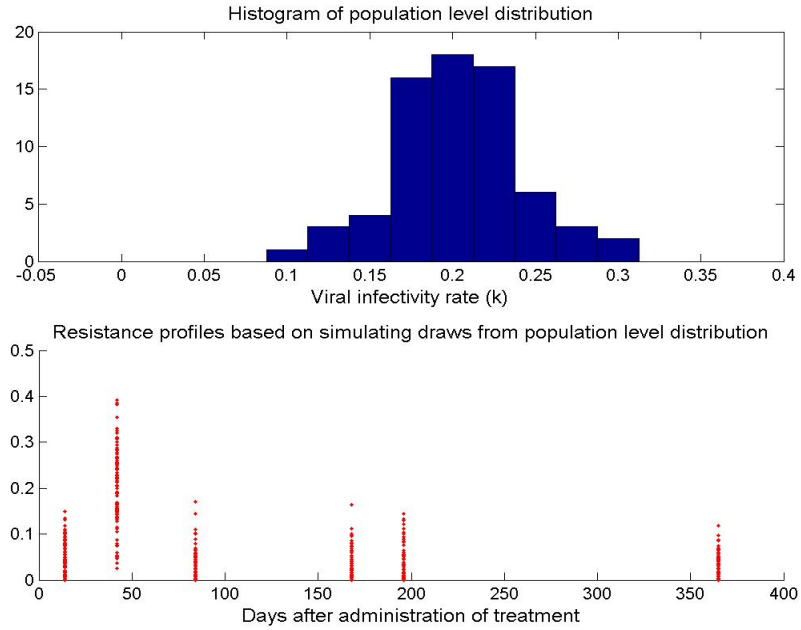

Fig. 5.2: Population level distribution for a hypothetical data set. A normal distribution with a mean 0.2 and a variance of 0.1 is assumed.

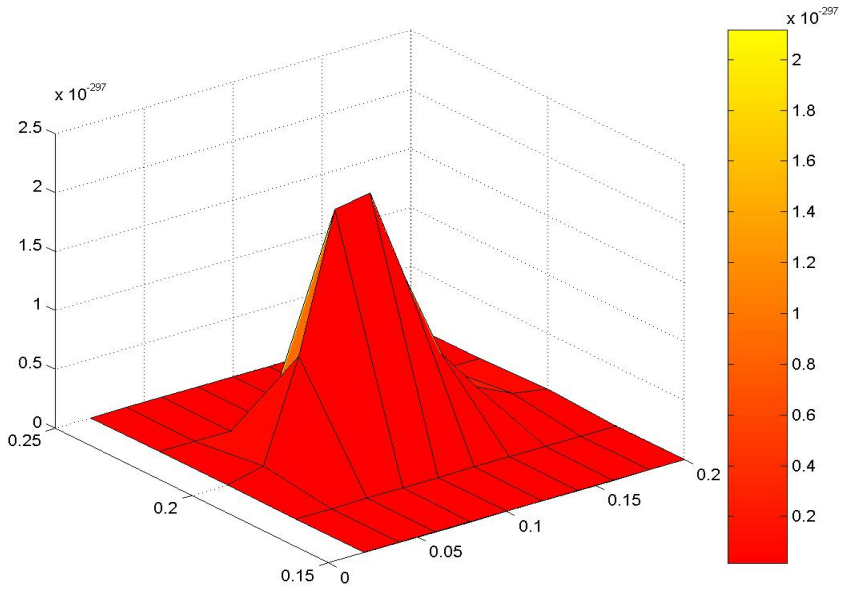

Fig. 5.3: Joint posterior probability density for mean and variance of population level distribution for treatment efficiency.

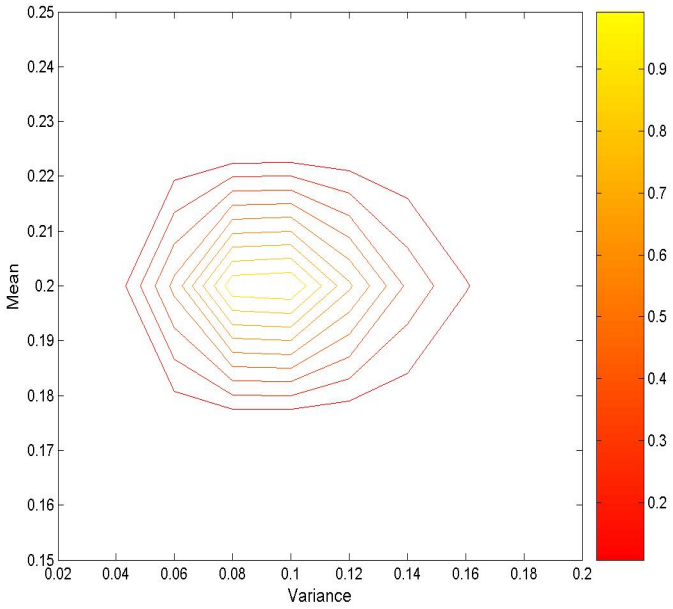

Fig. 5.4: Contour plot of joint posterior probability density for mean and variance of population level distribution for treatment efficiency. Contours represent points of equal posterior probability.

Recovering the population level distribution associated with a hypothetical and ideal data set is of course expected. Hypothetical data sets are useful to explore other properties of an inference scheme, apart from confirming their correct implementation. In particular, it is useful to estimate how many individual profiles are required to recover a known population level distribution. Repeating this analysis a few times while varying the number of individual profiles in each inference scheme, provided the following (rough) estimate of: 1) 40 profiles for estimating the mean of a normal population level distribution, and 2) 60 profiles for estimating both the mean and variance of a normal population level distribution. More nuanced aspects of hierarchical inference schemes can also be investigated: 1) to what extent does the inferred population distribution depend on how well it is sampled or represented by individual distributions. For example, 1) how many profiles must correspond to parameters drawn from the tail of the population level distribution, 2) what is the impact of noise in measured individual profiles on inferred population level parameters, and so on.

The rough estimates given above are sufficient to conclude that the data set used in this project (see Chapter 3) is unlikely to give accurate estimates of population level distribution. In fact, many profiles are incomplete, and the data set is characterized by infrequent sampling, which is possibly linked to deteriorating patient health. The data set had not been collected to meet the above mentioned inference

requirement. Having noted the limited usefulness of this data set for hierarchical Bayesian inference schemes, the next section investigates how the 'number of stages' can be estimated from a single complete profile.

5.4 Estimating survival distributions for populations of infected cells

Concerns are raised in [5] regarding the basis and methodology for estimating parameters for lifetime distributions, other than the usually assumed exponential lifetime distributions. The purpose of this section is to address these concerns by investigating a method for inferring other types of lifetime distributions. It is worth pointing out that inferring the lifetime distribution of infected cells is an ill-posed inverse problem: the discussions of Section 2.5 and Section 2.7 show that the dynamical behavior of age structured models depends in detail on the maternity distributions of infected cells. The problem is ill-posed because there is no unique solution; different maternity schedules can lead to the same dynamical behavior, depending on the relationship between the mean, variance and reproductive number of these schedules. We illustrate one way of adding prior information into the inverse problem.

We are going to assume that lifetime distributions of infected cells are accurately represented by gamma distributions. A gamma distribution is determined by a location and a shape parameter. When the location parameter, or the mean of the lifetime distribution, is kept constant, then no additional parameters are required when investigating gamma lifetime distributions. By fixing the mean value of the lifetime distribution a parameter relating to the shape of the distribution may be estimated from data instead. The next section investigates methods to estimate this parameter.

Homogeneous prior for age distribution

A homogeneous distribution is defined as a distribution that assigns equal probability to events of occupying equal volume in the sample space. For this simple model the sample space is the integers $(1, 2, \ldots, \infty)$. In particular, $n = 1$ corresponds to the assumption of an exponential lifetime distribution, $n \to \infty$ to the delta distribution, and $1 < n << \infty$ to a continuum of gamma distributions between these two extremes. All these distributions have the same mean and differ only in shape. In this sense, the sample space $(1, 2, \ldots, \infty)$ is an indirect way of sampling $(n_1, n_2, \ldots, n_\infty)$, the standard deviations of the lifetime distributions. The variance converges to zero as $n \to \infty$ when the lifetime distribution converges to a delta distribution, sharply peaked about the average life time. Consider Figure 5.5, where the shape of the lifetime distribution is characterized by the standard deviation of each lifetime distribution. It is clear that a sampling strategy giving equal probability to all stages $(n > 1)$, would be biased towards models corresponding to sharply peaked lifetime

distributions. The blue lines are an attempt to transform and categorize n into a new parameter n^*, over which a homogenous distribution can be defined.

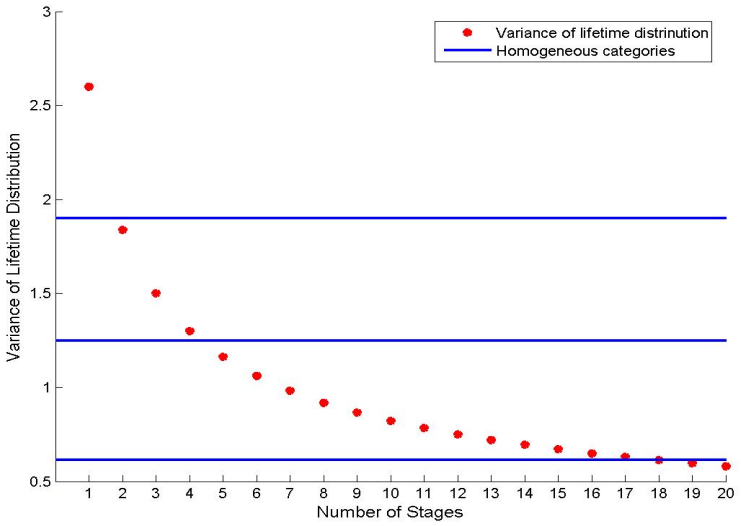

Fig. 5.5: Variance of lifetime distribution decreases as 'number of stages' increase.

The idea is to assign the same probability to all model parameters for which the corresponding variances lie between the same blue lines. The exception is n_1, the exponential distribution, and n_1 will simply be mapped to n_1^*. This transformation is listed below:

$$n_1^* = n_1 \tag{5.16}$$
$$n_2^* = n_2 \to n_4 \tag{5.17}$$
$$n_3^* = n_5 \to n_{18} \tag{5.18}$$
$$n_4^* = n_{19} \to \infty \tag{5.19}$$

There is no intrinsic reason why this is the correct way to introduce a homogenous distribution over the space $\{n_1^*, n_2^*, n_3^*, n_4^*\}$, and similar arguments for alternatives could be used. The following argument considers the relation between the tail of the lifetime distribution, and the rate a which the population will be eradicated during antiviral treatment. Consider a population with a constant mortality rate and no reproduction, i.e. consider m_1, and turn off the infectivity parameter from an arbitrary time after equilibrium had been reached. The half life, $\lambda^{\frac{1}{2}}$, of the population is defined to be the time when half of the population of infected cells are cleared. Recall that $l(a) = e^{-\int_0^a \mu(s)\,ds}$ represents the probability that an individual will survive from birth to age a, where $\mu(a)$ is the mortality rate experienced by the individual. We see that the half life is the time when

$$l(t) = e^{-\int_0^t \mu_P(t)\, dt} = \frac{1}{2} \tag{5.20}$$

For the exponential distribution, $\mu(t) = \mu_P$ is constant and the well known formula follows:

$$\lambda^{\frac{1}{2}} = \frac{log(2)}{\mu_P} \tag{5.21}$$

After 7 half lives less than 1% of the population remains. The choice of 1% is somewhat arbitrary, but corresponds approximately to the ratio before and after a period of treatment, during which the viral load drops from equilibrium to being undetectable. Figure 5.6 displays time to 99% clearance as the number of stages increase:

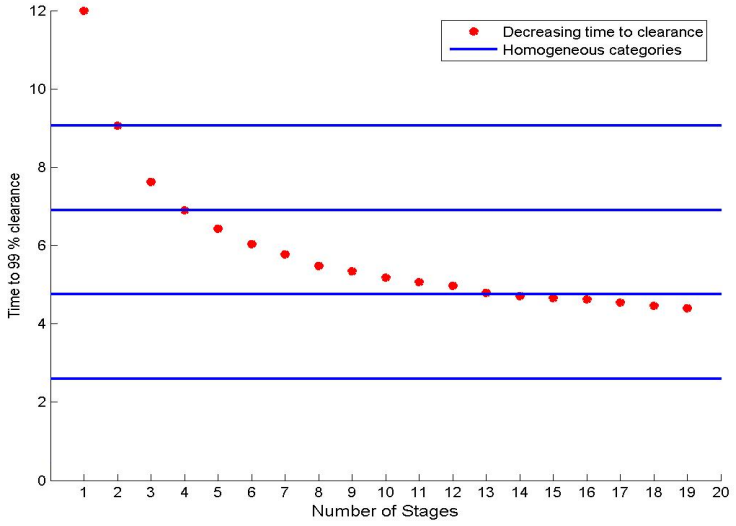

Fig. 5.6: Time to clearance decreases as 'number of stages' increase.

Note that the time to 99% percent clearance for M_1 is approximately $7 * \lambda^{\frac{1}{2}}$ for the exponential distribution. Also, that the time to clearance for $M_i, i \to \infty$ tend to the average life time of approximately 2.6 days. The corresponding homogeneous distribution is given by:

$$n_1^* = n_1 \tag{5.22}$$
$$n_2^* = n_2 \to n_4 \tag{5.23}$$
$$n_3^* = n_5 \to n_{13} \tag{5.24}$$
$$n_4^* = n_{14} \to \infty \tag{5.25}$$

Note that these two arguments lead to very similar homogeneous distributions for the 'number of stages' parameter. We adopt the homogenous prior given by equations (5.22)-(5.25) in the inference scheme. Estimating n^* has the following Bayesian solution:

$$p(n^*|d) = \frac{p(n^*)p(d|n^*)}{p(d)} \tag{5.26}$$

where,

- $p(n^*)$ is the prior for n^*. With no prior information on the distribution over n^*, it seems reasonable to assign equal probabilities to each n_i^*. Thus $p(n_i^*) = \frac{1}{4}$, which translates to:

$$p\left(n_i\right) = \begin{cases} \frac{1}{4} & : \quad i = 1 \\ (\frac{1}{3})(\frac{1}{4}) & : \quad i = 2, 3, 4 \\ (\frac{1}{9})(\frac{1}{4}) & : \quad i = 5, \ldots, 13 \\ \frac{1}{4} & : \quad i >= 14 \end{cases} \tag{5.27}$$

  and a sensible prior for a scale parameter σ_R is given by:

- The likelihood function, $p(d|n^*)$ is assumed to follow a Gaussian distribution:

$$p(d|n) = \prod_{i=1}^{I_D} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(d_i - g(n^*))^2}{2\sigma^2}} \tag{5.28}$$

- The marginal distribution $p(d)$ is given by:

$$p(d) = \sum_n p(n)p(d|n) \tag{5.29}$$

Data set d (Table 5.1) refers to the resistance profile of patient $R21$ (Appendix F). It consists of resistance measurements made on certain dates after treatment with a single dose of Nevirapine. The posterior distribution of the discrete parameter n, the number of stages, is displayed in Figure 5.7 and Figure 5.8. We have assumed treatment periods of 21 and 11 days for the situations depicted in Figure 5.7 and Figure 5.8 respectively.

Tab. 5.1: Resistance profile of patient $R21$

Patient	Dates	Resistance
$R21$	$42, 84, 196, 365$	$0.28, 0.34, 0.06, 0.01$

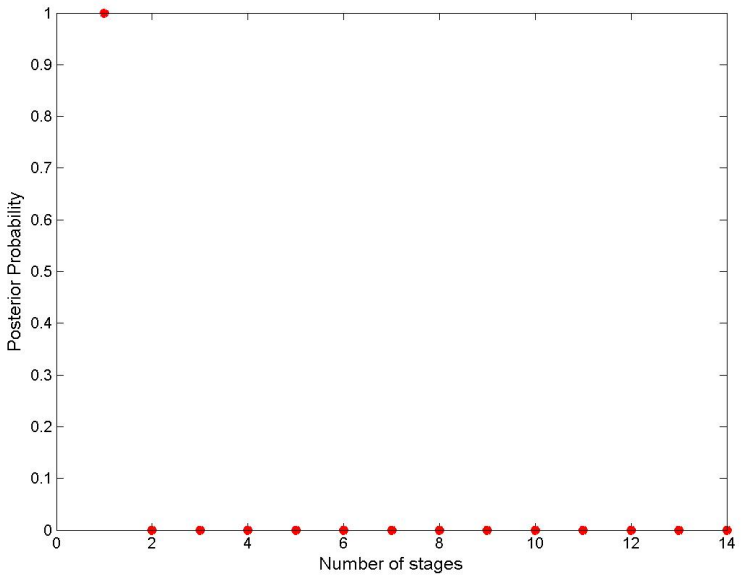

Fig. 5.7: An exponential lifetime distribution ($n = 1$ at point of maximum posterior distribution) is consistent with a treatment period of 21 days.

5.5 Summary

This chapter has applied the Bayesian framework as a sensible way to combine information for population dynamics. This information consists of: 1) forward models (developed in Chapter 2), 2) resistance profiles ('measured' in Chapter 3), and 3) prior distributions for parameters to be inferred. One of the strengths of Bayesian modelling is its ability to handle information that has a hierarchical structure. The first level of structure is built into the forward model; i.e. details of individual cells are structured into equations for populations of infected cells. Another level is built into a population level prior distribution. This level of modelling has been applied to a hypothetical data set, demonstrating that (roughly) 60 resistance profiles can shed light on the mean and variance of a population level parameter.

The main purpose of this chapter is to gain supporting evidence for the proposi-

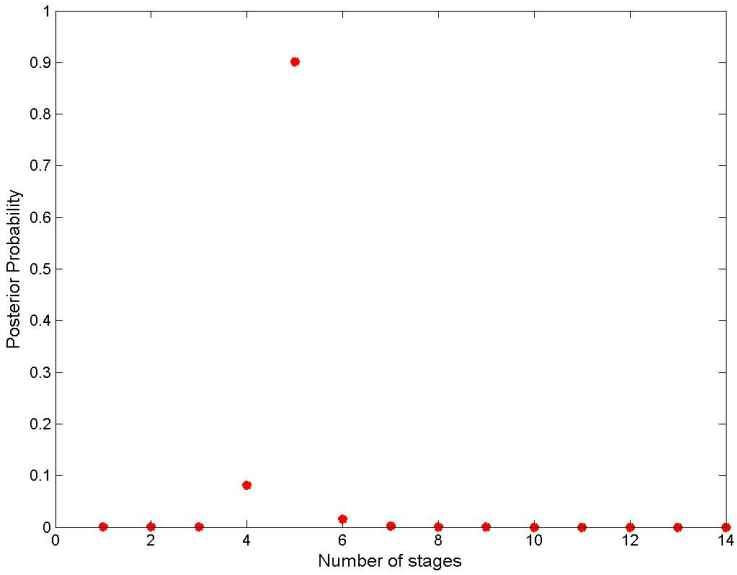

Fig. 5.8: A delta-like lifetime distribution ($n = 5$ at point of maximum posterior distribution) is consistent with a treatment period of 11 days.

tion that the lifetime distribution of infected cells may follow short-tailed distributions. A physiological model has been developed in Section 2.5, which demonstrates how short-tailed lifetime distributions can be obtained in a way that is insensitive to the exact details of the proposed physiological model. A mathematical forward model was built in Section 2.4 as a staged system of ODE's, where the staging parameter corresponds to n, the number of life stages. This parameter is also the shape parameter of gamma-like lifetime distributions. The shape of gamma-like distributions is not at odds with shapes that can be predicted by the physiological model of Section 2.5. Gamma-like lifetime distributions, via staged ODE's, provide a way of keeping the average lifetime fixed, while varying only the variance (essentially the shape of the tail) of the distribution. Another motivation for this modelling approach, apart from it simple implementation, is that allows us to have the commonly used theory as a special case. We have introduced a homogeneous prior (Figures 5.5 and 5.6) to reduce n to 4 categories, since the exact value is not important. The inference scheme is designed to merely distinguish the possibility of long-tailed, medium-tailed and short-tailed lifetime distributions. A value of 5 was inferred for the number of stages, assuming an effective treatment period of 11 days. This shape lies somewhere between an exponential and a delta distribution, and Figure 2.3 can be used to visualize its approximate shape.

A more conclusive inference result will require a more suitable data set. Such a result should ideally be based on the posterior distribution of a population level parameter, but our available data set is not sufficient for hierarchical Bayesian inference.

The data set made available for this project was obtained for substantially different purposes, not for our modelling and inference calculations. It has nevertheless demonstrated that a quantitative picture of strain competition can be obtained from PCR measurements, which can in principle be used to obtain a posterior estimate for the shape of the lifetime distribution of infected cells.

Chapter 6

Conclusions

In this project we have explored mathematical models for HIV-dynamics with a view to gaining a better understanding of the features of models with structured virion and infected cell populations. The simplest of models are usually not structured beyond the viral genotype and mayor cell phenotype level. These unstructured models have slow response times to environmental perturbations, for example, by anti-viral treatment. It was demonstrated that transient phenomena, such as the rapid emergence of a resistant strain, are realistically captured by including age structure for populations of infected cells. Age structure introduces considerable extra complexity into HIV dynamics models. It was shown that simplified age distributions, such as gamma distributions, can be modelled quite easily by introducing suitable categories of stages traversed by infected cells. Knowledge of physiological mechanisms at the binding sites on the surfaces of cells suggests constraints on how these models should depend on age or development in general. A simple combinatorial argument, which counts the number of viral peptides loaded onto the MHC complex of an infected cell, shows that infected cells can reasonably be expected to have short-tailed or gamma-like lifetime distributions. The Escalator Boxcar Train (EBT) algorithm was implemented to simulate a system of partial and integro-differential equations for two competing viral strains, where infected cells can have arbitrary age structure. Using this algorithm, it was demonstrated that strain competition depends not only on the mean age at which infected cells produce new virions and the average number that are produced, but on the entire maternity schedule.

Quantitative estimates of the relative fractions of $K103$ and $K103N$, the wild-type and Nevirapine resistant strains respectively, were obtained from a PCR experiment. A PCR processing toolbox was developed to process raw fluorescence data and to investigate the PCR process at a rudimentary level. The toolbox includes processing routines for: 1) normalizing the reporter signal with respect to the passive signal, 2) estimating and removing the background signal, 3) estimating the relative concentrations of viral subpopulations $103A$, 103_C, 103_G and 103_T, 4) correcting estimates for mispriming effects, based on a calibration matrix estimated from all calibration wells, and 5) collecting final estimates into resistance profiles for all trial attendees. A crude noise model was obtained as a repeatability statistic of the process.

A Bayesian inference framework was constructed for estimating viral parameters from the resistance data set. The solution of a Bayesian inverse problem (the posterior distribution) is the result of combining all available information. After stating a noise model for the measurement process, appropriate prior distributions were stated for the model parameters required to be inferred. An important notion in inverse problem theory is that of a homogenous, minimally informative or vague distribution. Shannon's measure of information then allows for the information content of a given distribution to be compared to that of a homogeneous distribution. The principle of translation invariance is used to derive a minimally informative prior distribution for location and scale parameters, such as the mean and variance of a normal distribution respectively. It was also demonstrated how population level priors can be used, whenever parameters can be viewed as independent draws from their joint population distribution. Population level priors are widely used in Bayesian interpretations of longitudinal data sets, where models must allow within-subject and between-subject variations.

The Bayesian inference framework was applied to estimate viral parameters from the Nevirapine resistance data set. It was found that simple and commonly used models for in-vivo HIV dynamics have slow response times, as it required almost three weeks of 'perfect' treatment to produce a significant drug resistant population. This conflicts with clinical evidence, which suggests a much faster emergence of resistant strains. A class of simple compartmental models for HIV dynamics was extended to include 'age' structure for populations of infected cells. It was demonstrated how resistance can be produced rapidly, by giving infected cells short-tailed survival distributions. A simple inference scheme was developed to infer the approximate shape of these survival distributions. The shape of this distribution was identified with the 'number of life stages' traversed by an infected cell, and mortality with the rate at which the last stage is approached. A homogenous prior was constructed for the 'number of stages' parameter, using two intuitive arguments. This was deemed necessary because it becomes increasingly difficult to distinguish the dynamical behavior of models structured according to an increasing number of life stages. This behavior converges to that of a fixed delay model, where cells all live up to a certain age after which they all 'die'. It was demonstrated that 1) the posterior distribution peaks at 1 stage when a treatment period of 21 days is assumed, and 2) peaks at 5 stages when a treatment period of 11 days is assumed. We have put forward supporting evidence for the claim that infected cells follow short-tail survival distributions, rather than evidence for a claim about an exact shape. Although we have developed an hierarchical approach to infer individual viral parameters, we do not have a suitable data set to infer population level viral parameters. We have demonstrated by means of inference on a hypothetical data set, that more than 60 individual viral load profiles are required to infer the mean and variance of a population level distribution. It would be useful to establish a relationship between accuracy in data obtained from the PCR process and the resolution at which population level parameters can be inferred.

Which aspects of this research project can be taken forward? A number of interesting questions have been asked in this project, which must be addressed to gain a better understanding of viral dynamics model and also of their interpretation from viral load data. These issues include:

- Is it possible to develop models for the cell life cycle that are more physiologically informative?
- Can effective population dynamics models be produced in a systematic way using scaling arguments?
- Substantial effort is required to quantify and control errors in PCR processes. Probe-based measurements may provide more accurate estimates of relative viral fractions, and this possibility is well worth exploring.

Appendix A

Viral life cycle

The following is a brief review of the main features, characterizing this life cycle into sequential steps [30, 6]:

- *Attachment and cell entry*: When a virion 'collides' with a target cell, it may attach itself, depending on local properties of the binding site. Affinity (or strength of the bond) is a function of the complementarity in shape and charge between the GP120 viral envelope protein (GP-spikes) and the receptors on the target cell membrane. If the area of interaction at the binding site is large enough to include a sufficient number of these GP-spike-and-receptor attachments, the viral and cellular membranes will fuse and virus empties its core (consisting of viral RNA and proteins) into cytoplasm of its host.

- *Reverse Transcription*: Once inside the cell the virus begins to exploit the cell's processes in order to synthesize new virus components. Using enzymes, the virus is able to initiate and optimize these processes for maximal production of viral protein sequences.

  The *Reverse Transcriptase* enzyme, of which multiple copies are present in the viral core, first reads the sequence of viral RNA nucleic acids and then transcribes this sequence into double stranded DNA. This transcription process is error-prone, producing single base mutations at every transcription cycle. Mutants that are more (genetically) distinct are produced at a rapidly declining rate.

  The import of this DNA complex into the host cell chromosome is signaled by viral proteins, followed by integration into the chromosome using the viral *Integrase* enzyme. Once integrated, the DNA is known as provirus.

- *Production* of viral proteins and genomes

  4) *Packaging and cell exit*: Viral proteins and RNA (genomes) are packaged into an envelope consisting of host cell membrane. The new virions are now equipped with envelope proteins, and exits the cell by 'budding' from the membrane, initiating new infections of susceptible cells.

Viral life cycles are more complex than this simplification, and each step in the process, comprise several steps. Viral production, for example, is a highly regulated

dynamic process, optimized with regard to resources required from the host cell and the advantages strategy to maximize viral production rate. It involves assembly of viral proteins and genomes, transport within the host cell, fusion with and release from host cell membrane.

Appendix B

Equilibrium state for multi strain dynamics

The equilibrium state for an n-strain model cannot be obtained by tedious algebraic manipulations of equations (2.11),(2.12) and (2.13). It can be obtained by means of a perturbation analysis, demonstrated here for a 2-strain model. A similar result is given by [6] but they use a more naive model, where mutation occurs only from the dominant to the sub-dominant strain. Consider the following 2-strain model:

$$\begin{aligned}
\frac{dT(t)}{dt} &= S_T - u_T T - k_1 V_1 T_1 - k_2 V_2 T && \text{(B.1)}\\
\frac{dP_1(t)}{dt} &= f k_1 V_1 T + \epsilon_{21} k_2 V_2 T - \mu_P P_1 && \text{(B.2)}\\
\frac{dP_2(t)}{dt} &= f k_2 V_2 T + \epsilon_{12} k_1 V_1 T - \mu_P P_2 && \text{(B.3)}\\
\frac{dV_1(t)}{dt} &= N_1 \mu_{P_1} P_1 - \mu_V V_1 && \text{(B.4)}\\
\frac{dV_2(t)}{dt} &= N_2 \mu_{P_2} P_2 - \mu_V V_2 && \text{(B.5)}
\end{aligned}$$

The meaning of all symbols is explained in Section 2.2. The state variables in the equation are $T(t)$, $P(t)$ and $V(t)$, representing total counts of healthy T cells,'productively infected' T cells and free virions, respectively. The system is put into its equilibrium state by setting the time derivatives of all the state variables to zero:

$$\begin{aligned}
0 &= S_T - u_T T - k_1 V_1 T_1 - k_2 V_2 T && \text{(B.6)}\\
0 &= f k_1 V_1 T + \epsilon k_2 V_2 T - \mu_P P_1 && \text{(B.7)}\\
0 &= f k_2 V_2 T + \epsilon k_1 V_1 T - \mu_P P_2 && \text{(B.8)}\\
0 &= N_1 \mu_{P_1} P_1 - \mu_V V_1 && \text{(B.9)}\\
0 &= N_2 \mu_{P_2} P_2 - \mu_V V_2 && \text{(B.10)}
\end{aligned}$$

where,

- We have assumed that strains $V_{1,2}$ have the same mutation rate.

- Mortality rates μ_P and μ_V are assumed to be equal for both classes of infected cells $P_{1,2}$ and viral strains $V_{1,2}$.
- We are going to assume that V_1 is the wild-type and V_2 the less fit strain. Differential fitness is captured by $k_2 < k_1$.
- Viral strains $V_{1,2}$ participate in short time scale dynamics and may be replaced in equations (B.6),(B.7),(B.8) by rescaling infectivity parameters $k_{1,2}$ such that $k_{1,2} \to \frac{k_{1,2}N_{1,2}\mu_{P_{1,2}}P_{1,2}}{\mu_V}$. The notation $X_{1,2}$ indicates a property associated with strain 1 and strain 2 respectively.

A perturbation analysis is set up to make small corrections to the equilibrium state of a model without mutation, a state which is easy to obtain. The feature of eventual extinction of the less fit strain can be used in an approximation scheme to estimate the equilibrium state for $P_{1,2}$. When there is no mutation, i.e. when $\epsilon = 0$, the less fit strain will receive no contributions from a mutating dominant strain. Thus for $\epsilon = 0$, we have the following solution:

$$P_2(0) = 0 \tag{B.11}$$
$$T(0) = \frac{\mu_P}{fk_1} \tag{B.12}$$
$$P_1(0) = \frac{S - \mu_T T}{k_1 T} \tag{B.13}$$

The system is expanded in terms of the small parameter ϵ:

$$T(\epsilon) = T(0) + \epsilon T^{(1)} + \epsilon^2 T^{(2)} + O(\epsilon^3) \tag{B.14}$$
$$P_1(\epsilon) = P_1(0) + \epsilon P_1^{(1)} + \epsilon^2 P_1^{(2)} + O(\epsilon^3) \tag{B.15}$$
$$P_2(\epsilon) = P_2(0) + \epsilon P_2^{(1)} + \epsilon^2 P_2^{(2)} + O(\epsilon^3) \tag{B.16}$$

where,

- $T(0), P_1(0), P_2(0)$ are the zero'th order ($\epsilon = 0$) solutions given by equations (B.11),(B.12),(B.13)
- $T^{(n)}, P_1^{(n)}, P_2^{(n)}$ indicate higher order terms.

The system can now be written in terms of these expansions:

$$0 = S - \mu_T T(\epsilon) - k_1 P_1(\epsilon) T(\epsilon) - k_2 P_2(\epsilon) T(\epsilon) \tag{B.17}$$
$$0 = fk_1 P_1(\epsilon) T(\epsilon) + \epsilon k_2 P_2(\epsilon) T(\epsilon) - \mu_P P_1(\epsilon) \tag{B.18}$$
$$0 = fk_2 P_2(\epsilon) T(\epsilon) + \epsilon k_1 P_1(\epsilon) T(\epsilon) - \mu_P P_2(\epsilon) \tag{B.19}$$

Collecting only first order terms, and omitting all terms involving $P_2(0) = 0$, we have the following $O(\epsilon)$ equations:

$$\begin{aligned} 0 &= \mu_T T^{(1)} - k_1 P_1^{(1)} T(0) - k_2 P_2^{(1)} T(0) && \text{(B.20)} \\ 0 &= f k_1 P_1^{(1)} T(0) + f k_1 P_1(0) T^{(1)} - \mu_P P_1^{(1)} && \text{(B.21)} \\ 0 &= f k_2 P_2^{(1)} T(0) + k_1 P_1(0) T(0) - \mu_P P_2^{(1)} && \text{(B.22)} \end{aligned}$$

which can be solved for the $O(\epsilon)$ terms:

$$\begin{aligned} T^{(1)} &= 0 && \text{(B.23)} \\ P_1^{(1)} &= -\frac{k_2}{k_1} P_2^{(1)} && \text{(B.24)} \\ P_2^{(1)} &= \frac{k_1 P_1(0) T(0)}{\mu_P - f k_2 T_0} && \text{(B.25)} \end{aligned}$$

The solutions of $P_{1,2}^{(1)}$ reflect the fact that first order corrections due to mutations are negative for P_1. Parameters μ_P, f, k_2 and $T(0)$ are constrained by the relationship $\mu_P - f k_2 T(0) > 0$, as first order corrections to P_2 must be positive, and by definition the total count of T cells infected with strain 2 will be increasing due to mutations of strain 1. The 'first order' equilibrium states are:

$$\begin{aligned} T_{eq} &= \frac{\mu_P}{f k_1} && \text{(B.26)} \\ P_{1_{eq}} &= \left(\frac{S - \mu_T T(0)}{k_1 T(0)} \right) - \epsilon \left(\frac{k_2}{k_1} P_2^{(1)} \right) + O(\epsilon^2) && \text{(B.27)} \\ P_{2_{eq}} &= \epsilon \left(\frac{k_1 P_1(0) T(0)}{\mu_P - f k_2 T_0} \right) + O(\epsilon^2) && \text{(B.28)} \end{aligned}$$

Appendix C

Equilibrium state for multi-strain dynamics including age structure

Appendix B derives the equilibrium state of a 2-strain model, assuming that the lifetime distributions of infected cell populations follow exponential distributions. A similar perturbation analysis can be used to calculate the equilibrium condition for general lifetime distributions. Consider the following 2-strain model, where general age-distributions are given to populations of infected cells. Restating equations (2.33)-(2.36):

$$\frac{dT(t)}{dt} = S_T(t) - T(t)\sum_{i=1}^{N_s} k_i V_i(t) - \mu_T T(t) \tag{C.1}$$

$$\frac{\partial P_i(a,t)}{\partial t} = -\frac{P_i(a,t)}{da} - \mu_{P_i}(a)P_i \tag{C.2}$$

$$P_i(0,t) = f k_i V_i(t) T_i(t) + T(t)\sum_{j\neq i}^{N_s} \varepsilon_{ji} k_j V_j(t) \tag{C.3}$$

$$\frac{dV_i(t)}{dt} = \int_0^\infty m_i(a) P_i(t)\, da - \mu_{V_i} V_i(t) \tag{C.4}$$

where,

- $i = 1, 2$ and $N_s = 2$.

The equilibrium state is obtained by setting the time derivatives of all the state variables to zero:

$$
\begin{aligned}
0 &= S_T - \mu_T T(t) - k_1 V_1(t) T(t) - k_2 V_2(t) T(t) && \text{(C.5)}\\
-\frac{\partial P_1(a,t)}{\partial a} &= -\mu_{P_1}(a) P_1(a,t) && \text{(C.6)}\\
-\frac{\partial P_2(a,t)}{\partial a} &= -\mu_{P_2}(a) P_2(a,t) && \text{(C.7)}\\
0 &= \int_0^\infty m_1(a) P_1(a)\, da - \mu_V V_1(t) && \text{(C.8)}\\
0 &= \int_0^\infty m_2(a) P_2(a)\, da - \mu_V V_2(t) && \text{(C.9)}
\end{aligned}
$$

where values for state variables refer to the values of the system when it is in equilibrium. In Section 2.3 it was shown that when a population $p(a,t)$ experiences time independent but age dependent mortality rates $\mu(a)$, then $p(a,t)$ obeys:

$$
\begin{aligned}
p(a,t) &= p(0,t-a) e^{-\int_0^a \mu(s)\, ds} && \text{(C.10)}\\
&= p(0,t-a) l(a) && \text{(C.11)}
\end{aligned}
$$

where $l(a) = e^{-\int_0^a \mu(s)\, ds}$ is the probability that an individual will survive from birth to age a. Using equation (C.11), we have the following boundary conditions:

$$
\begin{aligned}
P_1(a) &= \left(f k_1 V_1(t) T(t) + \epsilon k_2 V_2(t) T(t)\right) l_1(0,a) && \text{(C.12)}\\
P_2(a) &= \left(f k_2 V_2(t) T(t) + \epsilon k_1 V_1(t) T(t) l_2\right)(0,a) && \text{(C.13)}
\end{aligned}
$$

The system is again expanded in terms of the mutation rate ϵ, which is assumed to be a small parameter:

$$
\begin{aligned}
T(\epsilon) &= T(0) + \epsilon T^{(1)} + \epsilon^2 T^{(2)} + O(\epsilon^3) && \text{(C.14)}\\
P_1(\epsilon) &= P_1(0) + \epsilon P_1^{(1)} + \epsilon^2 P_1^{(2)} + O(\epsilon^3) && \text{(C.15)}\\
P_2(\epsilon) &= P_2(0) + \epsilon P_2^{(1)} + \epsilon^2 P_2^{(2)} + O(\epsilon^3) && \text{(C.16)}\\
V_1(\epsilon) &= V_1(0) + \epsilon V_1^{(1)} + \epsilon^2 V_1^{(2)} + O(\epsilon^3) && \text{(C.17)}\\
V_2(\epsilon) &= V_2(0) + \epsilon V_2^{(1)} + \epsilon^2 V_2^{(2)} + O(\epsilon^3) && \text{(C.18)}
\end{aligned}
$$

The zeroth order solution, i.e. setting $\epsilon = 0$, is obtained in Section 2.3:

$$
\begin{aligned}
T(0) &= \frac{\mu_V}{f k N_1} && \text{(C.19)}\\
P_1(0) &= \left(S_T f - \frac{\mu_T \mu_V}{k N_1}\right) l_1(0,a) && \text{(C.20)}\\
V_1(0) &= \frac{S_T f N_1}{\mu_V} - \frac{\mu_T}{k} && \text{(C.21)}\\
P_2(0) &= 0 && \text{(C.22)}\\
V_2(0) &= 0 && \text{(C.23)}
\end{aligned}
$$

The system can now be written in terms of ϵ expansions:

$$
\begin{aligned}
0 &= S - \mu_T T(\epsilon) - k_1 V_1(\epsilon) T(\epsilon) - k_2 V_2(\epsilon) T(\epsilon) && \text{(C.24)}\\
V_1(\epsilon) &= \frac{1}{\mu_V} \int_0^\infty m_1(a) \left[f k_1 V_1(\epsilon) T(\epsilon) + \epsilon k_2 V_2(\epsilon) T(\epsilon) \right] l_1(0,a)\, da && \text{(C.25)}\\
V_2(\epsilon) &= \frac{1}{\mu_V} \int_0^\infty m_2(a) \left[f k_2 V_2(\epsilon) T(\epsilon) + \epsilon k_1 V_1(\epsilon) T(\epsilon) \right] l_2(0,a)\, da && \text{(C.26)}\\
& && \text{(C.27)}
\end{aligned}
$$

Collecting only first order terms, and omitting all terms involving $P_2(0) = 0$, we have the following $O(\epsilon)$ equations for $T^{(1)}$, $V_1^{(1)}$ and $V_2^{(1)}$:

$$
\begin{aligned}
0 &= \mu_T T^{(1)} - k_1 V_1^{(1)} T(0) - k_2 V_2^{(1)} T(0) && \text{(C.28)}\\
V_1^{(1)} &= \frac{1}{\mu_V} \int_0^\infty m_1(a) f k_1 \left[V_1^{(1)} T(0) + V_1(0) T^{(1)} \right] l_1(0,a)\, da && \text{(C.29)}\\
V_2^{(1)} &= \frac{1}{\mu_V} \int_0^\infty m_2(a) f k_2 V_2^{(1)} T(0) l_2(0,a)\, da \\
&\quad + \frac{1}{\mu_V} \int_0^\infty m_2(a) k_1 V_1(0) T(0) l_2(0,a)\, da && \text{(C.30)}
\end{aligned}
$$

which can be solved for the $O(\epsilon)$ terms:

$$
\begin{aligned}
T^{(1)} &= 0 && \text{(C.31)}\\
V_1^{(1)} &= -\frac{k_2}{k_1} V_2^{(1)} && \text{(C.32)}\\
V_2^{(1)} &= \frac{k_1 N_2 V_1(0) T(0)}{\mu_V - f k_2 N_2 T(0)} && \text{(C.33)}
\end{aligned}
$$

The integrals on the left hand side of equations C.29 and C.30 simplify using $N_{1,2} = \int_0^\infty m_{1,2}(a) l_{1,2}(0,a)\, da$. Equilibrium states are given by:

$$
\begin{aligned}
T_{eq} &= \frac{\mu_V}{f k N_1} && \text{(C.34)}\\
V_{1eq} &= \left(\frac{S_T f N_1}{\mu_V} - \frac{\mu_T}{k} \right) - \epsilon \left(\frac{k_2}{k_1} V_2^{(1)} \right) + O(\epsilon^2) && \text{(C.35)}\\
V_{2eq} &= \epsilon \left(\frac{k_1 N_2 V_1(0) T(0)}{\mu_V - f k_2 N_2 T(0)} \right) && \text{(C.36)}\\
P_{1eq}(a) &= \left(f k_1 V_{1eq} T_{eq} + \epsilon k_2 V_{2eq} T_{eq} \right) l_1(0,a) + O(\epsilon^2) && \text{(C.37)}\\
P_{2eq}(a) &= \left(f k_2 V_{2eq} T_{eq} + \epsilon k_1 V_{1eq}(t) T_{eq} \right) l_2(0,a) + O(\epsilon^2) && \text{(C.38)}
\end{aligned}
$$

Appendix D

The escalator boxcar train algorithm

The EBT algorithm was introduced in Section 2.6 as a tool for numerically investigating the dynamics of structured populations. The algorithm follows cohorts of population members trough time and keeps track of the total population and age count of each cohort. A key requirement for successful implementation is that each cohort can be characterized by their average 'age'. The purpose of this section is to determine factors that contribute to the accuracy of the algorithm. This discussion follows [11] closely, but focuses on age, instead of more general physiological structure such as size, length, and so on. Consider an age structured population that evolves in time according to:

$$\frac{\partial p(a,t)}{\partial t} = -\frac{\partial p(a,t)}{\partial a} - \mu(a)p(a,t) \tag{D.1}$$

$$P(t) = \int_0^\infty p(a,t)da \tag{D.2}$$

$$p(0,t) = \int_0^\infty n(a)p(a,t)\,da \tag{D.3}$$

$$p(a,0) = p_0 \tag{D.4}$$

where,

- $p(a,t)$ is the population density and $P(t)$ the population count.
- $n(a)$ is the number of offspring born to an individual aged a.
- $p(0,t)$ and $p(t,0)$ are boundary and initial conditions respectively.

Consider next the total population $P(t)$, age count $A(t)$ and average age $\overline{a}$ of the population in a time interval (t_1, t_2):

$$P(t) = \int_{t_1}^{t_2} p(a,t)\,da \tag{D.5}$$

$$A(t) = \int_{t_1}^{t_2} ap(a,t)\,da \tag{D.6}$$

$$\overline{a} = \frac{\int_{t_1}^{t_2} ap(a,t)\,da}{\int_{t_1}^{t_2} p(a,t)\,da} \tag{D.7}$$

The derivative of $P(t)$ with respect to time is given by:

$$\begin{aligned}
\frac{dP(t)}{dt} &= \int_{t_1}^{t_2} \frac{\partial p(a,t)}{dt}\,da + p(t_2,t)\frac{d(t_2)}{dt} - p(t_1,t)\frac{d(t_1)}{dt} && \text{(D.8)}\\
&= \int_{t_1}^{t_2} \frac{\partial p(a,t)}{\partial t}\,da + \int_{t_1}^{t_2} \frac{\partial p(a,t)}{\partial a}\,da && \text{(D.9)}\\
&= -\int_{t_1}^{t_2} \mu(a)p(a,t)\,da && \text{(D.10)}
\end{aligned}$$

where the terms $\frac{d(t_2)}{dt} = 1$ and $\frac{d(t_2)}{dt} = 1$ make explicit reference to the fact that the integration range is a function of time. This reference will be omitted from the remainder of the analysis. A key part of the method is to obtain approximations to integrals of the type $\int_{t_1}^{t_2} \phi(a)p(a,t)\,da$, where $\phi(a)$ is a weighting function for $p(a,t)$. [11] obtains an approximation by Taylor expanding $\phi(a)$ around the value $a = \overline{a}$, where $\overline{a}$ is the average age in the interval (t_1, t_2). Expanding $\mu(a)$ around $a = \overline{a}$ gives the following approximation for $P(t)$:

$$\begin{aligned}
\frac{dP(t)}{dt} &= -\int_{t_1}^{t_2} \mu(\overline{a})p(a,t)\,da - \int_{t_1}^{t_2} \mu'(\overline{a})(a-\overline{a})p(a,t)\,da && \text{(D.11)}\\
&\quad -\frac{1}{2}\int_{t_1}^{t_2} \mu(\overline{a})''(a-\overline{a})^2 p(a,t)\,da\\
&= -\mu(\overline{a})\int_{t_1}^{t_2} p(a,t)\,da - \mu'(\overline{a})\int_{t_1}^{t_2} (a-\overline{a})p(a,t)\,da && \text{(D.12)}\\
&\quad -\mu''(\overline{a})\frac{1}{2}\int_{t_1}^{t_2} (a-\overline{a})^2 p(a,t)\,da
\end{aligned}$$

The term involving μ' is zero, as substituting the definition of $\overline{a}$ given by equation (D.7) shows. The term involving $\mu''(a-\overline{a})^2$ corresponds to how much age varies in the cohort, and can be omitted if the variation is small. Omitting all terms in the expansion involving $\frac{d^n\mu}{da}$ where $n \geq 2$ gives:

$$\begin{aligned}
\frac{dP(t)}{dt} &= -\mu(\overline{a})\int_{t_1}^{t_2} p(a,t)\,da && \text{(D.13)}\\
\frac{dP(t)}{dt} &= -\mu(\overline{a})P(t) && \text{(D.14)}
\end{aligned}$$

and hence $P(t)$ can be tracked by an ODE. Deriving an ODE for $A(t)$ in the time interval (t_1, t_2) proceeds along the same lines:

$$\frac{dA(t)}{dt} = \int_{t_1}^{t_2} \frac{d}{dt}\left(ap(a,t)\right)\, da \tag{D.15}$$

$$= \int_{t_1}^{t_2} p(a,t)\, da + \int_{t_1}^{t_2} a\frac{d}{dt}p(a,t)da \tag{D.16}$$

$$= \int_{t_1}^{t_2} p(a,t)\, da - \int_{t_1}^{t_2} a\mu(a)p(a,t)\, da \tag{D.17}$$

If second and higher order corrections are omitted, as they were for the simplification of equation (D.13) hold, then

$$\frac{dA(t)}{dt} = P(t) - \mu(\overline{a})A(t) \tag{D.18}$$

Equations (D.14) and (D.18) hold for cohorts born at time 0, i.e $0 < t_1 < t_2$. A different expansion is used for the boundary cohort (i.e. the cohort into which new individuals are recruited), as the average age will be ill-defined when in p(a,t) is zero equation (D.7). This issue is resolved by expanding around t_b, instead of around the average age $\overline{a}$. Another approach is to simply ignore this potential problem [3], and to use the expansion given by equations (D.14) and (D.18) for all cohorts.

Consider the total population $P_0(t)$ and age count $A_0(t)$ of the initial cohort in a time interval (t_b, t_1), where t_b is the time at birth:

$$\frac{dP_0(t)}{dt} = \int_{t_b}^{t_1} p(a,t)\, da \tag{D.19}$$

$$A_0(t) = \int_{t_b}^{t_1} (a - t_b)p(a,t)\, da \tag{D.20}$$

The time derivative of $P_0(t)$ is given by:

$$\frac{dP_0(t)}{dt} = \frac{d}{dt}\int_{t_b}^{t_1} p(a,t)\, da \tag{D.21}$$

$$= \int_{t_b}^{t_1} \frac{\partial p(a,t)}{dt}\, da + p(t_1,t) \tag{D.22}$$

noting that the lower bound of integration t_b is independent time, and the term $-p(t_b,t)$ is therefore not present in the right hand side of equation (D.22). It can however be added, together with $+p(t_b,t)$, to obtain:

$$\frac{dP_0}{dt}(t) = \int_{t_b}^{t_1} \frac{\partial p(a,t)}{\partial t}\, da + \int_{t_b}^{t_1} \frac{\partial p(a,t)}{\partial a}\, da + p(t_b,t) \tag{D.23}$$

$$= -\int_{t_b}^{t_1} \mu(a)p(a,t)\, da + \int_{t_b}^{t_1} n(a)p(a,t)\, da \tag{D.24}$$

where the second term on the left hand side of the equation (D.24) is boundary condition D.3. Equation (D.24) is approximated by expanding $\mu(a)$ around the value $a = t_b$:

$$\begin{aligned}
\frac{dP_0(t)}{dt} &= -\int_{t_b}^{t_1} \mu(t_b)p(a,t)\,da - \int_{t_b}^{t_1} \mu'(t_b)(a-t_b)p(a,t)\,da && \text{(D.25)}\\
&\quad -\frac{1}{2}\int_{t_b}^{t_1} \mu''(\overline{a})(a-t_b)^2 p(a,t)\,da + \int_{t_b}^{t_1} n(a)p(a,t)\,da \\
&= -\mu(t_b)P(t) - \mu'(t_b)A_0(t) + \int_{t_b}^{t_1} n(a)p(a,t)\,da && \text{(D.26)}
\end{aligned}$$

omitting all terms in the expansion involving $\frac{d^n\mu}{da}$ where $n \geq 2$. An approximation for $A_0(t)$ is obtained in a similar way:

$$\begin{aligned}
\frac{dA_0(t)}{dt} &= \int_{t_b}^{t_1} p(a,t)\,da + \frac{d}{dt}\int_{t_b}^{t_1} (a-t_b)p(a,t)\,da && \text{(D.27)}\\
&= P_0(t) - \int_{t_b}^{t_1} (a-t_b)\frac{\partial p(a,t)}{\partial t}\,da + (t_1-t_b)p(t_1,t) && \text{(D.28)}\\
&= P_0(t) - \int_{t_b}^{t_1} (a-t_b)\mu(a)p(a,t)\,da + (t_1-t_b)p(t_b,t) && \text{(D.29)}\\
&= P_0(t) - \int_{t_b}^{t_1} (a-t_b)\mu(a)p(a,t)\,da && \text{(D.30)}\\
&\quad +(t_1-t_b)\int_{t_b}^{t_1} n(a)p(a,t)\,da
\end{aligned}$$

expanding $\mu(a)$ around $a = t_b$ gives:

$$\frac{dA_0(t)}{dt} = P_0(t) - \mu(t_b)A_0(t) + (t_1-t_b)\int_{t_b}^{t_1} n(a)p(a,t)\,da \qquad \text{(D.31)}$$

To summarize, we have obtained the following ODE's to track cohorts through time:

$$\begin{aligned}
\frac{dP(t)}{dt} &= -\mu(\overline{a})P(t) && \text{(D.32)}\\
\frac{dA(t)}{dt} &= P(t) - \mu(\overline{a})A(t) && \text{(D.33)}
\end{aligned}$$

while the boundary or birth cohort is tracked by:

$$\begin{aligned}
\frac{dP(t)}{dt} &= -\mu(t_b)P(t) - \mu'(t_b)A_0(t) + \int_{t_b}^{t_1} n(a)p(a,t)\,da && \text{(D.34)}\\
\frac{dA_0(t)}{dt} &= P_0(t) - \mu(t_b)A_0(t) + (t_1-t_b)\int_{t_b}^{t_1} n(a)p(a,t)\,da && \text{(D.35)}
\end{aligned}$$

These equations were all obtained by truncating an expansion in age dependent mortality $\mu(a)$ at second or higher order terms. This truncation is justified if, in each cohort, mortality varies linearly with age. These ODE's will therefore track cohorts accurately when the mortality can be approximated by piecewise linear functions, where the domains of these functions are given by the cohort boundaries. The assumption of linearly varying mortality can always be attained by decreasing the size of the age cohorts. This increased resolution will naturally be accompanied by increased computational effort.

D.1 Matlab code

This section lists the Matlab (Version 7.0) code used in Section 2.7 to explore strain competition. The function *MATERNITY-SCEDULES.m* can be used to generate the age-structures for two competing strains depicted in Figure 2.6. The lifetime distributions are gamma distributions and they are not implicitly related to maternity functions $m(a)$. The function creates age structures for strain 1 and strain 2, with the following properties:

- They have the same mean age.
- They were also chosen to have the same burst size, namely 500 virions born to infected cells of either type.
- They differ however in the variance of the age at which new virions are produced: strain 2 has a higher variance in the age at giving birth to new virions.

The functions required for producing the two age structures are: 1) *MATERNITY-SCHEDULES.m*, 2) *GAMMA-LIFE.m* and 3) *EXP-PRODUCTION.m*.

The function *ESCALATOR-BOX-CAR.m* is the starting point for an implementation of the 'Escalator Box Car' algorithm applied to modelling competition between two strains. In this model strain 2 starts with an initial count of zero, and is created from mutating strain 1. The slight fitness advantage of strain 2 allows it to replace and dominate strain 1 eventually, even though strain 1 has won the race to dominance during primary infection. The functions required for running the 'Escalator Box Car' algorithm are: 1) *ESCALATOR-BOX-CAR.m*, 2) *INTEGRATE-MODEL.m*, 3) *EVALUATE-DERIVATIVES.m*, 4) *sEVALUATE-DERIVATIVES.m*, 5) *INTERP-CURVE.m* and 6) *da-dmu.m*.

D.1.1 Hypothetical Age structures for two competing strains

```
function MATERNITY_SCHEDULES(k1,k2,n1,n2,a1,a2,b1,b2)
%
% This function calls GAMMA_LIFE to calculate age structures
% (survivorship function, lifetime distribution, maternity
% or production function, and age dependent mortality) of an
% infected cell. Results are displayed at the end.
%
% INPUT:
```

```
% k1 - Number of stages for strain1.
% k2 - Number of stages for strain2.
% n1 - The max number of virions produced by a cell infected with strain1.
% n2 - The max number of virions produced by a cell infected with strain2.
% a1 - Age at which strain1 starts producing virions.
% a2 - Age at which strain2 starts producing virions.
% b1 - Rate at which strain1 approached max production rate.
% b2 - Rate at which strain2 approached max production rate.
%
% OUTPUT:
% The age structures for the two strains are saved in
% MAT_SCHEDULE.mat. It is loaded in the BOX-CAR algorithm.
%
% Example of typical use
% [age,m,L,dl,mu,G]=MATERNITY_SCHEDULE(6,3,370,308,1,0.7,0.4,0.5)

age=0:0.25:10;
avg_age=2.6; %days
[Ages1,Maternity1,Survivorship1,d_Survivorship1,Mortality1,LT1]=GAMMA_LIFE(age,k1,(1/avg_age),a1,b1,n1);
'Total number of virions produced by cell infected with strain1:'
N1=sum(Maternity1(:,2).*Survivorship1(:,2))

mat1=Maternity1(:,2).*Survivorship1(:,2); mat1=mat1/N1;
age_mat1=Maternity1(:,1);

mt1=trapz(mat1.*age_mat1); 'Variance of maternity schedule of cell
infected with strain1:'
mv1=trapz(mat1.*((age_mat1-mt1).*(age_mat1-mt1)))

[Ages2,Maternity2,Survivorship2,d_Survivorship2,Mortality2,LT2]=GAMMA_LIFE(age,k2,(1/avg_age),a2,b2,n2);
'Total number of virions produced by cell infected with strain2:'
N2=sum(Maternity2(:,2).*Survivorship2(:,2))

mat2=Maternity2(:,2).*Survivorship2(:,2); mat2=mat2/N2;
age_mat2=Maternity2(:,1);

'Variance of maternity schedule of cell infected with strain2:'
mt2=trapz(mat2.*age_mat2);

mv2=trapz(mat2.*((age_mat2-mt2).*(age_mat2-mt2)))

%//////////////////////////////////////////////////////////////////////////
%plot age structures
close all;

h1=figure(1); subplot(2,1,1); hold on;
h_l=plot(Survivorship1(:,1),LT1,'r'); set(h_l,'linewidth',2);
h_l=plot(Survivorship1(:,1),Survivorship1(:,2),'b-');
set(h_l,'linewidth',2);
h_l=plot(Maternity1(:,1),Maternity1(:,2)/max(Maternity1(:,2)),'g-');
set(h_l,'linewidth',2);
h_l=plot(Mortality1(:,1),Mortality1(:,2),'k-');
set(h_l,'linewidth',2);
legend('L(a)','l(0,a)','m(a)','u(a)','Location','NorthEast');
title('Age structure for strain 1'); xlabel('Days');

figure(h1); subplot(2,1,2); hold on;
h_l=plot(Survivorship2(:,1),LT2,'r'); set(h_l,'linewidth',2);
h_l=plot(Survivorship2(:,1),Survivorship2(:,2),'b-');
set(h_l,'linewidth',2);
h_l=plot(Maternity2(:,1),Maternity2(:,2)/max(Maternity2(:,2)),'g-');
set(h_l,'linewidth',2);
h_l=plot(Mortality2(:,1),Mortality2(:,2),'k-');
set(h_l,'linewidth',2);
legend('L(a)','l(0,a)','m(a)','u(a)','Location','NorthEast')
title('Age structure for strain 2'); xlabel('Days');

%//////////////////////////////////////////////////////////////////////////
%plot maternity schedules
h2=figure(2); subplot(2,1,1); hold on
h_l=plot(Maternity1(:,1),mat1,'g-'); set(h_l,'linewidth',2);
h_l=plot([mt1 mt1],[0 1],'k:'); set(h_l,'linewidth',2);
h_l=plot([mt1+mv1 mt1+mv1],[0 1],'k:'); set(h_l,'linewidth',2);
h_l=plot([mt1-mv1 mt1-mv1],[0 1],'k:'); set(h_l,'linewidth',2);
axis([0 6 0 0.3]); title('Maternity distribution for strain
1','FontSize',[12]); xlabel('Days','FontSize',[12]);

figure(h2) subplot(2,1,2); hold on

h_l=plot(Maternity2(:,1),mat2,'g-'); set(h_l,'linewidth',2);
h_l=plot([mt2 mt2],[0 1],'k:'); set(h_l,'linewidth',2);
h_l=plot([mt2+mv2 mt2+mv2],[0 1],'k:'); set(h_l,'linewidth',2);
h_l=plot([mt2-mv2 mt2-mv2],[0 1],'k:'); set(h_l,'linewidth',2);
```

```
axis([0 6 0 0.3]); title('Maternity distribution for strain
2','FontSize',[12]); xlabel('Days','FontSize',[12]);

%//////////////////////////////////////////////////////////////////////
%save the age structures for use in the BOX-CAR algorithm

MAT_SCHEDULE{1,1}=Ages1; MAT_SCHEDULE{2,1}=Maternity1;
MAT_SCHEDULE{3,1}=Survivorship1;
MAT_SCHEDULE{4,1}=d_Survivorship1; MAT_SCHEDULE{5,1}=Mortality1;
MAT_SCHEDULE{6,1}=LT1;

MAT_SCHEDULE{1,2}=Ages2; MAT_SCHEDULE{2,2}=Maternity2;
MAT_SCHEDULE{3,2}=Survivorship2;
MAT_SCHEDULE{4,2}=d_Survivorship2; MAT_SCHEDULE{5,2}=Mortality2;
MAT_SCHEDULE{6,2}=LT2;

save MAT_SCHEDULE.mat MAT_SCHEDULE

end %function MATERNITY_SCHEDULES

function [age,m,L,dL,mu,G]=GAMMA_LIFE(age,k,muP,a_m,b,P_max)
%
% This function calculates the survivorship function, lifetime distribution,
% maternity or production function, and age dependent mortality of an
% infected cell
%
% The lifetime distributions are gamma distributions and they
% are not implicitly related to maternity functions.
%
% INPUT:
% age - Array containing age discritization (in days), ex. age=[0:0.25:10]
% k - The number of stages traversed by an infected cell.
% muP - Mortality rate.
% a_p - The earliest age at which the onset of viral production begins.
% b - Controls the rate at which the maximum production rate is approached.
% P_max - The max number of virions produced by an infected cell.
%
% OUTPUT:
% age - Array containing age discritization (in days), ex. age=[0:0.25:10]
% m - Virion production schedule.
% L - Survivorship function.
% dL - Derivative of L.
% mu - Mortality rate.
% G - Gamma lifetime distribution.
%
% Example of typical use
% [age,m,L,dl,mu,G]=GAMMA_LIFE_D(1,1,0:0.25:10,5,0.3846,370,1,0.4000)
%

lambda= muP*k; for j=1:size(age,2)

    % Calc gamma distribution
    % gamma is a matlab function
    G(j)=((lambda^k)*(age(j)^(k-1))*(exp(-1*lambda*age(j))))/(gamma(k));

    % Calc gamma cumulative distribution function
    % gamma cdf is a matlab function
    G2(j)=gamcdf(age(j),k-1,1/lambda);
    G3(j)=G(j)/(1-G2(j));

    m(j,1)= age(j);
    m(j,2)= EXP_PRODUCTION(age(j),a_m,b,P_max);

end

LT=ones(size(G2))-G2;

%Calc mu
mu(:,2)=diff(LT,1); for i=1:length(mu)
    mu(i,1) = age(i);
    mu(i,2)= (-1*mu(i,2))/(LT(i));
end

%Calc DL
dL(:,2)=-1*diff(LT); for i=1:length(dL)
    dL(i,1) = age(i);
    dL(i,2)= dL(i,2);
end

%Calc L
L(:,1)=age; L(:,2)=LT;

end %function GAMMA_LIFE
```

```
function p = EXP_PRODUCTION(cell_age,a_p,b,P_max)
%
% This function calculates the number of virions produced by an infected
% cell using the exponential virion production schedule of equation (2.52).
%
% INPUT:
% cell_age - Age of infected cell.
% a_p - The earliest age at which the onset of viral production begins.
% b - Controls the rate at which the maximum production rate is approached.
% P_max - The max number of virions produced by an infected cell.
%
% OUTPUT:
% p - The number of virions produced at cell_age.
%

if(cell_age>=a_p)
    p = P_max*(1 - exp(-b*(cell_age-a_p)) );
else
    p = 0;
end

end %function EXP_PRODUCTION
```

D.1.2 An implementation of the 'Escalator Box Car' Algorithm

```
function ESCALATOR_BOX_CAR
%
% This function is where the 'escalator box car' algorithm starts.
% It is assumed that age structures for two competing strains have
% already been produced by means of MAT_SCHEDULES.m, which must be run first,
% using for example: MATERNITY_SCHEDULES(6,3,370,308,1,0.7,0.4,0.5)

% This function manages the population cohorts and the time intervals over
% which to integrate, calling the function INTEGRATE_MODEL.
%
% The function implements the algorithm explained in Section 2.6 and
% Appendix D. Matlab supports the parsing of string expressions into
% numerical values. This allows for great flexibility in implementing
% this algorithm. The time derivatives are first written as string
% expressions, containing terms for state and other variables. These string
% expressions are then parsed using the current values of all variables
% used in the expression, resulting in a vector containing the numerical
% values of the time derivatives, one value for each ODE in equations (2.56)
% and (2.58). The state and auxiliary variables are declared as global variables
% in order for their values to be continuously available in the functions:
% ESCALATOR_BOX_CAR->INTEGRATE_MODEL->EVALUATE_DERIVATIVES->sEVALUATE_DERIVATIVES

% Results are plotted at the end. Note that the function takes a while t0
% complete. Integrating the model detailed in this function, which ranges
% from day 0 to day 250 (set in variable T_MAX), takes +- 40 minutes.

%clear all variables. Important as global variables are used.
clear all;

%//////////////////////////////////////////////////////////
%Treatment start, duration and end
global RTIstart RTItime RTIend
%Initial T cell population
global T0
%T cell production rate
global S
%Life span assumptions, mortality rates
global muT muP1 muP2 muV
%Infectivity parameters
global k1nat k2nat
%Treatment efficiency
global teff1 teff2
%Probability of correct transcription
global f
%Number of virrions produced by infected cell during life span
global N1 N2
%Mutate_ji - transcription error required for strain j to infect cell with
%genome of strain i
global mutate_12 mutate_21

%Some model parameters can be adjusted here
PAR(1)=14;%RTItime
PAR(2)=0.1;%teff1
```

```
PAR(3)=0.9;%teff2
PAR(4)=0.00003; %mutate_12, mutate_21

%//////////////////////////////////////////////////////////////////////
%Init model
RTIstart=150; RTItime=PAR(1); RTIend=RTIstart+RTItime; muT=0.002;
muV=3;
%muP1,muP2 from MAT_SCHEDULE
T0=90000000000; S=muT * T0; k1nat=0.000000000004;
k2nat=0.000000000004; teff1=PAR(2); teff2=PAR(3); f=1;
%Ni from MAT_SCHEDULE
mutate_12=PAR(4); mutate_21=PAR(4);

%//////////////////////////////////////////////////////////////////////
%Age structures loaded from MAT_SCHEDULE.m
load MAT_SCHEDULE age=0:0.5:7; global Ages1 Maternity1
Survivorship1 d_Survivorship1 Mortality1 d_Mortality1 global Ages2
Maternity2 Survivorship2 d_Survivorship2 Mortality2 d_Mortality2

global Ages1 Maternity1 Survivorship1 d_Survivorship1 Mortality1
d_Mortality1 global Ages2 Maternity2 Survivorship2 d_Survivorship2
Mortality2 d_Mortality2

Ages1=MAT_SCHEDULE{1,1}; Maternity1=MAT_SCHEDULE{2,1};
Survivorship1=MAT_SCHEDULE{3,1};
d_Survivorship1=MAT_SCHEDULE{4,1}; Mortality1=MAT_SCHEDULE{5,1};
LT1=MAT_SCHEDULE{6,1};

Ages2=MAT_SCHEDULE{1,2}; Maternity2=MAT_SCHEDULE{2,2};
Survivorship2=MAT_SCHEDULE{3,2};
d_Survivorship2=MAT_SCHEDULE{4,2}; Mortality2=MAT_SCHEDULE{5,2};
LT2=MAT_SCHEDULE{6,2};

d_Mortality1 = da_dmu(Mortality1); d_Mortality2 =
da_dmu(Mortality2);

%////////////////////////////////////////////////
%EBT housekeeping variables
global nCohorts;%number of cohorts
nCohorts=length(age);
%Total cohort counts
global P_P1 P_P2 P_P1=zeros(1,nCohorts); P_P2=zeros(1,nCohorts);
%Total age in cohorts
global A_P1 A_P2 A_P1=zeros(1,nCohorts); A_P2=zeros(1,nCohorts);
%Average age in cohorts
global a_P1 a_P2 a_P1=zeros(1,nCohorts); a_P2=zeros(1,nCohorts);

T_MAX=10;%time to where the model must be calculated
T_STEP=age(2)-age(1);%age discritization
T_PAR=0:T_STEP:T_MAX;%time discritization
TPAR=cell(1);%declare a variable of type cell

TPAR{1}=T_PAR(1); INITIAL=0; MODEL=[]; m_c=0; for
i=1:length(T_PAR)-1

TPAR{2}=T_PAR(i+1)
[t,y,indT,indP1,indA1,indP2,indA2,indV1,indV2,INIT]=INTEGRATE_MODEL(TPAR,INITIAL);

%//////////////////////////////////////////////////////////////////////
%Load integration results to state variables
T_P = INIT(indT);

P_P1=INIT(indP1); A_P1=INIT(indA1); P_P2=INIT(indP2);
A_P2=INIT(indA2);

V_P1=INIT(indV1); V_P2=INIT(indV2);

%Save integration results to model
m_c=m_c+1; MODEL(:,m_c)=INIT';

%Move cohorts along the 'escalator'
for j=length(P_P1):-1:2
    P_P1(j)=P_P1(j-1);
    A_P1(j)=A_P1(j-1);
    P_P2(j)=P_P2(j-1);
    A_P2(j)=A_P2(j-1);
end

%Clear initial cohort
P_P1(1)=0;A_P1(1)=0;P_P2(1)=0;A_P2(1)=0;

%////////////////////////////////////////////////////////////////
%Use integration results as intitial values for new integration
```

```
sI='[T_P'; for j=1:nCohorts
    sp=sprintf('P_P1(%i)',j);
    sI=[sI,',',sp];
end %for j

for j=1:nCohorts
    sp=sprintf('A_P1(%i)',j);
    sI=[sI,',',sp];
end  %for j

for j=1:nCohorts
    sp=sprintf('P_P2(%i)',j);
    sI=[sI,',',sp];
end %for j

for j=1:nCohorts
    sp=sprintf('A_P2(%i)',j);
    sI=[sI,',',sp];
end %for j

sI=[sI,',','V_P1',',','V_P2',']'];

INITIAL=eval(sI);

%//////////////////////////////////////////////////////////////////////////
TPAR{1}=T_PAR(i+1);

end %for i

%Calc the total number of the two strains across all ages
for i=1:size(MODEL,2)
    D(i)=T_PAR(i);%value of time interval
    T=MODEL(indT,i);
    P1(i)=sum(MODEL(indP1,i));%total amount of strain1
    P2(i)=sum(MODEL(indP2,i));%total amount of strain2
end

'MODEL' size(MODEL)

close all; h1=figure; hold on;


%Display on a log-plot
h_l=plot([RTIstart RTIstart],[0 30],'k:'); set(h_l,'linewidth',2);
h_l=plot (D, log(P1),'r-'); set(h_l,'linewidth',2); h_l=plot (D,
log(P2),'g-'); set(h_l,'linewidth',2);

ylabel('Viral Load','FontSize',[12])
xlabel('Days','FontSize',[12])
%axis([0 max(D) 0 30]);

Ha = gca; legend(Ha,'off'); h_lr=plot([0],[0],'r.-');
h_lg=plot([0],[0],'g.-'); lg=legend([h_lr h_lg],'strain 1','strain
2','Location','NorthEast');


end %function ESCALATOR_BOX_CAR

function
[t,y,indT,indP1,indA1,indP2,indA2,indV1,indV2,INIT]=INTEGRATE_MODEL(TPAR,INITIAL)
%
% This function performs the integration for a time interval of the EBT algorithm explained in
% Section 2.6 and Appendix D. It is called from ESCALATOR_BOX_CAR.
%
% INPUT:
% TPAR - Specifies integration interval.
% INITIAL - Specifies values for state variables at integration start.
%
% OUTPUT:
% t - Time points of integration.
% y - Vector of state variables at integration time points.
% indT,indP1,indA1,indP2,indA2,indV1,indV2 - Indices to state variables.
% INIT - y(:,end), the final values for state variables to use as initial
%        value for next integration.
%

%Integration start and end
Tstart=TPAR{1}; Tend=TPAR{2};

%EBT housekeeping variables
global P_P1 P_P2 global A_P1 A_P2 global a_P1 a_p2

%Age structures
```

```
global Ages1 Maternity1 Survivorship1 d_Survivorship1 Mortality1
d_Mortality1 global Ages2 Maternity2 Survivorship2 d_Survivorship2
Mortality2 d_Mortality2 global nCohorts

%Initialize state variables for first integration step
if(INITIAL==0)

    %Initialize T
    global T0
    T = T0;

    P1=zeros(1,nCohorts);
    P1(1)=1;%one cell infected with strain1
    A1=zeros(1,nCohorts);
    P2=zeros(1,nCohorts);
    A2=zeros(1,nCohorts);

    P_P1=P1;
    A_P1=A1;
    P_P2=P2;
    A_P2=A2;

    %Initialize viral populations
    V1 =10;
    V2 =0;

    sI='[T';
    for i=1:nCohorts
    sp=sprintf('P1(%i)',i);
    sI=[sI,',',sp];
    end

    for i=1:nCohorts
    sp=sprintf('A1(%i)',i);
    sI=[sI,',',sp];
    end

    for i=1:nCohorts
    sp=sprintf('P2(%i)',i);
    sI=[sI,',',sp];
    end

    for i=1:nCohorts
    sp=sprintf('A2(%i)',i);
    sI=[sI,',',sp];
    end

    sI=[sI,',','V1',',',',V2',']'];

    % Integrates the system of differential equations y' = f(t,y) from time
    % Tstart to Tend
    [t y] = ode113(@EVALUATE_DERIVATIVES, [Tstart Tend], eval(sI),[],Tstart,Tend);
    % other ode solvers:    ode45, ode23, ode15s, ode23s, ode23t, ode23tb

    P1=P_P1;
    A1=A_P1;

else

    [t y] = ode113(@EVALUATE_DERIVATIVES, [Tstart Tend], INITIAL,[],Tstart,Tend);
    % other ode solvers: ode45, ode23, ode15s, ode23s, ode23t, ode23tb

end % if INITIALIZE MODEL


%Return final values to serve as initial values for next integration
INIT=y(end,:);

%Indices for state variables
indT =1; p1=2;indP1=[p1:p1+nCohorts-1];
a1=p1+nCohorts;indA1=[a1:a1+nCohorts-1];
p2=a1+nCohorts;indP2=[p2:p2+nCohorts-1];
a2=p2+nCohorts;indA2=[a2:a2+nCohorts-1]; indV1=indA2(end)+1;
indV2=indV1+1;

end %function INTEGRATE_MODEL

function  dydt=EVALUATE_DERIVATIVES(t,y,t_l,t_u)
%
% This function returns the derivatives, i.e. the right hand side of a
% system of ODE's, when requested by the integration solver.
%
% INPUT:
```

```
% t - Time points of integration.
% y - Vector of state variables at integration time points.
% t_l,t_u - Lower and upper bound of integration.
%
% OUTPUT:
% dydt - derivatives of state variables.
%

global RTIstart RTItime RTIend  muT muP1 muP2 muV T0 S global
k1nat k2nat teff1 teff2 N1 N2 f mutate_12 mutate_21 global
nCohorts

k1eff=k1nat; if((t>RTIstart)&(t<
RTIend)),k1eff=k1nat*(1-teff1);end k2eff=k2nat;
if((t>RTIstart)&(t< RTIend)),k2eff=k2nat*(1-teff2);end

%Load state variables
T=y(1); p1=1; for i=1:nCohorts
    sp=sprintf('P1(%i)=y(%i);',i,p1+i);
    eval(sp)
end %for i

a1=p1+nCohorts; for i=1:nCohorts
    sp=sprintf('A1(%i)=y(%i);',i,a1+i);
    eval(sp)
end %for i

p2=a1+nCohorts; for i=1:nCohorts
    sp=sprintf('P2(%i)=y(%i);',i,p2+i);
    eval(sp)
end %for i

a2=p2+nCohorts; for i=1:nCohorts
    sp=sprintf('A2(%i)=y(%i);',i,a2+i);
    eval(sp)
end %for i


p3=a2+nCohorts+1; V1=y(p3); V2=y(p3+1);

dydt = eval(sEVALUATE_DERIVATIVES(t,t_l,t_u))';

end %function EVALUATE_DERIVATIVES

function eq=sEVALUATE_DERIVATIVES(t,t_l,t_u)
%
% This function will return the string defining the derivatives, i.e. the right
% hand side of a system of ODE's for the implementation of the EBT algorithm to strain
% dynamics:
%
% T' = S - k1eff*T*V1- k2eff*T*V2 -muT
%
% P1,P2 is defined by the EBT explained in
% Section 2.6 and Appendix D.
%
% V1' = m1(a(1))*P1(a(1)) + .... + m1(a(n))*P1(a(n)) - muV * V1
% V2' = m2(a(1))*P2(a(1)) + .... + m2(a(n))*P2(a(n)) - muV * V2
%
% INPUT:
% t - Time points of integration
% t_l,t_u - Lower and upper bound of integration.
%
% OUTPUT:
% eq - String expression for the equations of the derivatives of the state
%      variables.
%


%EBT housekeeping variables
global P_P1 P_P2 global A_P1 A_P2 global a_P1 a_p2

%Age structures
global Ages1 Maternity1 Survivorship1 d_Survivorship1 Mortality1
d_Mortality1 global Ages2 Maternity2 Survivorship2 d_Survivorship2
Mortality2 d_Mortality2 global nCohorts

%Calculate average age in cohorts
global nCohorts a_P1=zeros(size(A_P1)); a_P2=zeros(size(A_P2));
da=t-floor(t); d_age1=Ages1(2)-Ages1(1); a_P1(1)=da; a_P2(1)=da;
for i=2:nCohorts

    if(P_P1(i)<=0)
        a_P1(i)=0;
```

```
    else
        a_P1(i)=A_P1(i)/P_P1(i);
    end

    if(P_P2(i)<=0)
        a_P2(i)=0;
    else
        a_P2(i)=A_P2(i)/P_P2(i);
    end

    if(a_P1(i)<0),a_P1(i)=0;end
    if(a_P2(i)<0),a_P2(i)=0;end

end % for i

%New/Birth cohort
cohort_b=1; b_age=0;

%Healthy (target cells)
eqT = 'S - k1eff*T*V1- k2eff*T*V2 - muT*T';

%Productively infected with strain1
eqP1=''; for n=1:length(P_P1)

eqP1n=''; if(n==cohort_b)
    age=b_age;
    mu=INTERP_CURVE(Mortality1,age);
    d_mu=INTERP_CURVE(d_Mortality1,age);
    eqP1n=sprintf('(f*k1eff*V1*T + mutate_21*k2eff*T*V2) - (%f)*A1(%i)  - (%f)*P1(%i)',d_mu,cohort_b,mu,cohort_b);
else
 age=a_P1(n);
     mu=INTERP_CURVE(Mortality1,age);
     eqP1n=sprintf('- (%f)*P1(%i)',mu,n);
end

if(length(eqP1)==0)
    eqP1=[eqP1n];
else
    eqP1=[eqP1,',',eqP1n];
end

end %for n

%Total age in cohorts of cells productively infected with strain1
eqA1=''; for n=1:length(A_P1)

if(n==cohort_b)
    age=b_age;
    mu=INTERP_CURVE(Mortality1,age);
    eqA1n=sprintf(' (%f)*(f*k1eff*V1*T + mutate_21*k2eff*T*V2)+ P1(%i) - (%f)*A1(%i)',(t-t_l),n,mu,n);
else
    age=a_P1(n);
    mu=INTERP_CURVE(Mortality1,age);
    eqA1n=sprintf(' P1(%i) - (%f)*A1(%i)',n,mu,n);
end

if(length(eqA1)==0)
    eqA1=[eqA1n];
else
    eqA1=[eqA1,',',eqA1n];
end

end %for n

%Productively infected with strain2
eqP2=''; for n=1:length(P_P2)

eqP2n=''; if(n==cohort_b)
    age=b_age;
    mu=INTERP_CURVE(Mortality2,age);
    d_mu=INTERP_CURVE(d_Mortality2,age);
    eqP2n=sprintf('f*k2eff*V2*T + mutate_12*k1eff*T*V1 - (%f)*A2(%i) - (%f)*P2(%i)',d_mu,cohort_b,mu,cohort_b);
else
    age=a_P2(n);
    mu=INTERP_CURVE(Mortality2,age);
    eqP2n=sprintf('- (%f)*P2(%i)',mu,n);
end

if(length(eqP2)==0)
    eqP2=[eqP2n];
else
    eqP2=[eqP2,',',eqP2n];
end
```

```
end %for n

%Total age in cohorts of cells productively infected with strain2
eqA2=''; for n=1:length(A_P2)

if(n==cohort_b)
    age=b_age;
    mu=INTERP_CURVE(Mortality2,age);
    eqA2n=sprintf(' (%f)*(f*k2eff*V2*T + mutate_12*k1eff*T*V1) + P2(%i) - (%f)*A2(%i)',(t-t_l),n,mu,n);
else
    age=a_P2(n);
    mu=INTERP_CURVE(Mortality2,age);
    eqA2n=sprintf(' P2(%i) - (%f)*A2(%i)',n,mu,n);
end

if(length(eqA2)==0)
    eqA2=[eqA2n];
else
    eqA2=[eqA2,',',eqA2n];
end

end %for n

%Strain1
eqV1= '0'; for n=1:length(P_P1)
    age=a_P1(n);
    m=INTERP_CURVE(Maternity1,age);
    eqV1n=sprintf(' (%f)*P1(%i) ',m,n);
    eqV1=[eqV1,'+',eqV1n];
end %for n
eqV1=[eqV1,'- muV*V1'];

%Strain2
eqV2= '0'; for n=1:length(P_P2)
    age=a_P2(n);
    m=INTERP_CURVE(Maternity2,age);
    eqV2n=sprintf(' (%f)*P2(%i) ',m,n);
    eqV2=[eqV2,'+',eqV2n];
end %for n

eqV2=[eqV2,'- muV*V2'];

%When evaluated using current values of state variables, this string
%expression will be an array containing derivatives
eq=['[',eqT,',',eqP1,',',eqA1,',',eqP2,',',eqA2,',',eqV1,',',eqV2,']'];

end %function sEVALUATE_DERIVATIVES

function v=INTERP_CURVE(curve,age)
%
% This function interpolates a value on 'curve' at particular 'age'.
%
v=0; X=curve(:,1); Y=curve(:,2);
v = interp1(X,Y,age);%matlab interpolate function

end %function INTERP_CURVE

function d_mu = da_dmu(mortality)
%This function calculates the derivative of mortality
dd_mu=diff(mortality(:,2)); for i=1:length(dd_mu)
    d_mu(i,1)=mortality(i,1);
    d_mu(i,2)=dd_mu(i);
end

end %function da_dmu
```

Appendix E

Linear models for age-dependent population dynamics

Models for structured population dynamics aim to describe how the distribution of a population changes over time, with respect to structure variables. Structure variables include age, size, mass and other variables that are correlated to individual development. This section investigates classical linear models for age-structured populations. These models can be used to describe population dynamics in constant or changing environments. If a population is subjected to constant environmental conditions, i.e maintains the same fertility and mortality schedules (vital rates), then a stable age distribution eventually emerges. Environmental changes perturb the vital rates, resulting in transient population dynamics.

The McKendrick-Von Foster equation is a partial differential equation for continuous age and time variables:

$$\frac{n(a,t)}{\partial t} = -\frac{n(a,t)}{\partial a} - \mu(a,t)n(a,t) \tag{E.1}$$

The *birth law* enters as a boundary condition for a:

$$n(0,t) = \int_0^\infty m(a)n(a,t)\,da,\ t > 0 \tag{E.2}$$

where,

- $m(a)$, the maternity function, is the expected number of offspring per individual aged a, in next time interval $(t, t+dt)$.
- The range of integration is determined by the earliest and latest age at which an individual can be productive. For simplicity it is taken to be 0 and ∞ respectively.
- $n(0,t) = B(t)$ may be viewed as birth rate at time t.

An initial age distribution serves as a boundary condition for t:

$$n(a,0) \quad = \quad n_0(a),\ a >= 0 \tag{E.3}$$

(E.1),(E.2) and (E.3) define the classic linear model of age-dependent population dynamics [46]. This system of equations is solved using the method of characteristics. Note that equation (E.1) can be written as

$$\frac{dn(a,t)}{dt} \quad = \quad -u(a)n(a,t) \tag{E.4}$$

along curves in the $A \times T$ plane for which $\frac{da}{dt} = 1$, i.e. along the lines $a - t = c$. The value of $n(a,t)$ can be obtained from any previous value $n(a-\alpha, t-\alpha)$, that is on the line $a - t = c$ through (a,t) and $(a-\beta, t-\beta)$:

$$n(a,t) \quad = \quad n(a-\beta, t-\beta)e^{-\int_{a-\beta}^{a} \mu(s)\,ds} \tag{E.5}$$

The general solution is obtained by extending this line until it intersects either the line $a = 0$ or $t = 0$, for $a < t$ and $a >= t$ respectively. The solution to equation (E.4) is then given by [46]:

$$n(a,t) \quad = \quad n(0, t-a)e^{-\int_{0}^{a} \mu(s)\,ds}, a < t \tag{E.6}$$

$$n(a,t) \quad = \quad n(a-t, 0)e^{-\int_{a-t}^{a} \mu(s)\,ds}, a \geq t \tag{E.7}$$

The solution has a natural interpretation in terms of survivorship. The probability of surviving the time interval between age a_1 and age a_2, where $0 <= a_1 <= a_2$, is given by:

$$l(a_1, a_2) = \frac{l(0,a_2)}{l(0,a_1)} \quad = \quad e^{-\int_{a_1}^{a_2} \mu(s)\,ds} \tag{E.8}$$

The probability of surviving from birth to age a will be abbreviated by l(a). In terms of the survival function, (E.8), (E.6) reduces to:

$$n(a,t) \quad = \quad n(0, t-a)l(0,a), a < t \tag{E.9}$$

$$n(a,t) \quad = \quad n(a-t, 0)l(a-t, a), a \geq t \tag{E.10}$$

where (E.9) represents those individuals born $t - a$ time units ago who survived to age a and (E.10) those members aged $a - t$ at time $t = 0$, who survived to age a. The population density function $n(a,t)$ would be completely determined if the birth rate $B(t)$ was known. Substituting $n(0,t)$ into (E.2) gives the birth rate [46]:

$$B(t) = \int_0^t m(a)n(a,t)\,da + \int_t^\infty m(a)n(a,t)\,da \tag{E.11}$$

$$= \int_0^t m(a)B(t-a)l(a)\,da + \int_t^\infty m(a)n(a-t,0)l(a-t,a)\,da \tag{E.12}$$

It is seen that birth rate $B(t)$ satisfies the renewal equation

$$B(t) = \int_0^\infty \widetilde{m}(a)B(t-a)\,da + B_0 \tag{E.13}$$

where,

- $\widetilde{m}(a) = m(a)l(a)$, the maternity schedule, is the 'kernel' of the integral (E.13).
- B_0 is the contribution to the birth rate from the population that existed at time $t = 0$. For simplicity it may be assumed that $B_0 = 0$, i.e that initial population given by the second integral in (E.13) has passed away.

Exponential population growth

Lotka studied a special class of solutions to (E.13), known as stable age distributions. They have the following form, obtained by separating variables a and t [46]:

$$n(a,t) = \mathbf{T}(t)\mathbf{A}(a) \tag{E.14}$$

Lotka made the assumption that the population will eventually grow exponentially. This is a reasonable assumption: a single self-generating population will eventually grow exponentially or die off exponentially. Thus, in the long time limit we have a natural basis in which to express the population density:

$$n(a,t) = e^{rt}\mathbf{A}(a) \tag{E.15}$$

Substituting (E.15) into (E.1) gives:

$$\frac{d\mathbf{A}}{da} = -\left(\mu(a) + r\right)\mathbf{A}(a) \tag{E.16}$$

with solution

$$\mathbf{A}(a) = \mathbf{A}(0)e^{-ra-\int_0^a \mu(a)\,da} \tag{E.17}$$

Boundary condition (E.2) implies:

$$1 \quad = \quad \int_0^\infty e^{-ra} m(a) l(a)\, da = g(r) \tag{E.18}$$

known as a *characteristic equation* for the asymptotic growth rate r. It has an infinite number of roots for a general maternity schedule $m(a)l(a)$ [2]. The asymptotic growth rate is determined by the unique positive real root, where uniqueness follows from the fact that $g(r)$ is a monotonically decreasing function of r. The number

$$R_0 = g(0) = \int_0^\infty m(a) l(a)\, da \tag{E.19}$$

is called the *reproductive value* of an individual. It is the expected number of offspring of the individual during its lifetime and it is a critical parameter:

- When $R_0 > 1$, the population grows exponentially.
- When $R_0 < 1$, the population dies out exponentially.
- When $R_0 = 1$, the population fluctuates stochastically.

The birth rate $B(t)$ can be written as [47]:

$$B(t) \quad = \quad \sum_1^\infty c_i e^{r_i t} \tag{E.20}$$

where $e^{r_i t}$ are solutions of the homogenous renewal equation:

$$B(t) \quad = \quad \int_0^\infty \widetilde{m}(a) B(t-a)\, da \tag{E.21}$$

It can be shown that a dominant eigenvalue exists, that its real part is larger than the real part of any other eigenvalue, and that it determines the asymptotic or long-time behavior of the population age-structure [47, 48]. Note that in, equation (E.14) we have selected this dominant eigenvalue $r = r_d$ by making the ansatz that the population time and stable age distribution separates in the long time limit according to this equation. It then follows that the population density will tend asymptotically to:

$$n(a,t) \sim e^{r_d t} A(0) e^{-r_d a} l(a) \tag{E.22}$$

where,

- Time dependence is given by $e^{r_d t}$.

- Age dependence is given by $A(0)e^{-r_d a}l(a)$.
- The *shape* of the stable age distribution tends to $e^{-r_d a}l(a)$.

This stable age distribution will emerge for each possible initial age distribution, i.e. the population will eventually 'forget' its initial conditions and grow or decay according to r_d.

Uniform change in mortality rate

The stable age distribution of a population governed by equation (E.18) is unaffected by a uniform change in mortality. This is demonstrated by [2], using the following simple approach. Suppose that mortality changes from $\mu(a)$ to $\mu(a)+c$. Survivorship $l(a)$ changes by a factor e^{-ca}:

$$l^*(a) = e^{-\int_0^a (\mu(a)-c)\,da} = e^{-ca}l(a) \tag{E.23}$$

Survivorship will decrease rapidly towards older ages if $c > 0$. The shape of the age distribution is given by:

$$e^{-(r_d-c)a}l^*(a) = e^{-(r_d-c)a}e^{-ca}l(a) = e^{-r_d a}l(a) \tag{E.24}$$

showing that the shape of the stable age distribution remains unchanged by a uniform change in mortality.

Dynamical behavior in response to changing environment

Lotka's model has shown how the stable age distribution can be obtained from the birth rate and the survivorship function, i.e. from the maternity schedule. It shows how a stable or characteristic age distribution emerges when stable environmental conditions prevail, i.e when the maternity schedule remains constant. The model exhibits transient population dynamics if environmental conditions change. The dynamical behavior of the population can be described in great detail by eigenvalue analysis on set of eigenvalues of characteristic equation (E.18).

An alternative approach derives from the fundamental relationship between the intrinsic rate of increase and moments of the maternity schedule, discussed in detail by [2]. It follows from the observation that the characteristic equation (E.18) is the Laplace transform of the maternity schedule $m(a)l(0,a)$:

$$g(r) \quad = \quad \int_0^\infty e^{-ra}m(a)l(a)\,da \tag{E.25}$$

Normalizing the maternity schedule with respect to R_0:

$$\frac{g(r)}{R_0} = \int_0^\infty e^{-ra} \frac{m(a)l(a)}{R_0}\, da \tag{E.26}$$

and taking the logarithm of $\frac{g(r)}{R_0}$ generates the cumulants [40, 1]:

$$\log\left(\frac{g(r)}{R_0}\right) = -log(R_0) = -r\kappa_1 + \frac{r^2\kappa_2}{2!} - \frac{r^3\kappa_3}{3!} + \ldots \tag{E.27}$$

or in terms of moments about the mean:

$$-\log(R_0) = -r\mu + \frac{r^2\sigma^2}{2!} - \ldots \tag{E.28}$$

omitting moments of higher order than σ^2, the variance of the normalized fertility schedule.

Equation (E.28) makes it possible to compare the intrinsic growth rate of populations with different maternity schedules, or to compare the intrinsic growth rate of the same population before and after a perturbation to its maternity schedule. Perturbation analysis gives the following equations for the derivatives of r with respect to σ^2, μ and R_0 [2]:

$$\frac{dr}{d\sigma^2} = \frac{r^2}{2(\mu - r\sigma^2)} \tag{E.29}$$

$$\frac{dr}{d\mu} = \frac{-r}{\mu - r\sigma^2} \tag{E.30}$$

$$\frac{dr}{dR_0} = \frac{1}{R_0(u - r\sigma^2)} \tag{E.31}$$

Equations (E.29)-(E.31) suggest the following comparison for the intrinsic rate of two populations, equal in all respects, except for a difference in μ, σ or R respectively:

- The population with a more dispersed age of birth will have a higher intrinsic growth rate.
- The population with a larger average age of birth will have a lower intrinsic growth rate. Convergence to the stable age distribution is more rapid when the average fertility is localized at young ages.
- The population with a larger reproductive value R_0 will have a higher intrinsic growth rate.

These relationships allow modellers to construct classes of $\{\mu, \sigma, R_0\}$ that would lead to equivalent changes in the natural rate of increase. These arguments may suggest functional forms for maternity schedules for populations of infected cells.

Appendix F

Nevirapine resistance data set

Listed on next page.

Tab. F.1: Nevirapine resistance data set.

Patient	Dates	Resistance
$R196$	$14, 42, 84, 168, 196, 365$	$0.01, 0.27, 0.22, 0.11, 0.02, 0.00$
$R108$	$42, 84, 196, 365$	$0.01, 0.00, 0.00, 0.00$
$R149$	$42, 84, 196, 365$	$0.32, 0.24, 0.00, 0.00$
$R16$	$42, 84, 196, 365$	$0.33, 0.08, 0.00, 0.00$
$R21$	$42, 84, 196, 365$	$0.28, 0.34, 0.06, 0.01$
$R24$	$42, 84, 196, 365$	$0.00, 0.00, 0.00, 0.00$
$R32$	$42, 84, 196, 365$	$0.03, 0.19, 0.11, 0.00$
$R35$	$42, 84, 196, 365$	$0.34, 0.06, 0.01, 0.00$
$R46$	$42, 84, 196, 365$	$0.02, 0.00, 0.00, 0.00$
$R70$	$42, 84, 196, 365$	$0.05, 0.00, 0.00, 0.00$
$D2$	$42, 42, 365$	$0.06, 0.00, 0.00$
$R116$	$42, 84, 196$	$0.17, 0.01, 0.00$
$R150$	$42, 196, 365$	$0.97, 0.16, 0.05$
$R17$	$84, 196, 365$	$0.02, 0.00, 0.00$
$R189$	$14, 196, 365$	$0.01, 0.00, 0.00$
$R191$	$42, 196, 365$	$0.12, 0.00, 0.00$
$R33$	$42, 196, 365$	$0.15, 0.00, 0.00$
$R88$	$42, 84, 365$	$0.27, 0.13, 0.01$
$R97$	$42, 84, 196$	$0.11, 0.00, 0.00$
$D108$	$42, 365$	$0.79, 0.00$
$D11$	$42, 365$	$0.02, 0.00$
$D150$	$42, 365$	$0.36, 0.00$
$D185$	$42, 365$	$0.65, 0.00$
$D188$	$42, 365$	$0.21, 0.00$
$D196$	$42, 365$	$0.18, 0.02$
$D20$	$42, 365$	$0.02, 0.00$
$D29$	$42, 365$	$0.20, 0.00$
$D32$	$42, 365$	$0.01, 0.00$
$D58$	$42, 365$	$1.00, 0.00$
$D70$	$42, 365$	$0.03, 0.00$
$R27$	$42, 365$	$0.19, 0.00$
$R58$	$84, 365$	$0.18, 0.00$
$R58$	$84, 365$	$0.18, 0.00$
$R8$	$196, 365$	$0.19, 0.00$
$D113$	42	0.09
$D115$	42	0.03
$D136$	42	0.19
$D149$	42	0.21
$D16$	365	0.04
$D164$	365	0.00
$D164B$	42	0.01
$D186$	365	0.00
$D191$	42	0.16

Appendix G

Solving inverse problems by combining states of information

A formulation for the solution of inverse problems as a combination of information has been introduced by [13], and this section provides a short review of this approach. Consider two probability measures P_1 and P_2, and let p_1 and p_2 be their probability densities respectively. A 'combination' of p_1 and p_2, given a homogenous distribution $\mu(x)$, is given by:

$$\Phi(x) = C(p_1, p_2) = \frac{p_1(x)p_2(x)}{\mu(x)} \tag{G.1}$$

It has the desired properties that:

- 1) $C(p_1, p_2)$ is invariant under reparametrization.
- 2) $C(p_1, p_2)$ is commutative; i.e. $C(p_2, p_1)$.
- 4) If $p_1(A) = 0$ for all events $A \subset X$, then combining p_1 with any P_2 gives zero.
- 5) $C(p, \mu) = p$; i.e. combining any distribution p with the homogeneous distribution does not change the information content of p.

It is demonstrated in [13, 49] that Φ is the only distribution that would 'combine' P_1 and P_2 in a way that is consistent with these properties. The following relation combines all information associated with any inverse problem, to obtain a posteriori information:

$$\Phi(d, m) = \frac{\rho(d, m)\theta(d, m)}{\mu(d, m)} \tag{G.2}$$

where,

- $\rho(d, m)$ is the joint probability density function for prior information, which could be reduced to

  $$\rho(d, m) \quad = \quad \rho_D(d)\rho_M(m) \tag{G.3}$$

  if a priori information of model parameters has been obtained independently of measurement.

- $\theta(d, m)$ represents theoretical information. When an explicit theory is available, it has the form:

$$\theta(d, m) \quad = \quad \theta(d|m)\mu_M(m) \tag{G.4}$$

  It is stated as a conditional distribution for d, given a homogeneous distribution over m.

  It can easily be shown that (G.2) reduces to Bayes's theorem when an explicit theory $\theta(d|m)$ is available. This can however not be done for a large class of problems, where the joint probability density $p(d, m)$ can only be specified in terms of accumulating joint observations in $D \times M$, the data and model space.

  Another feature of (G.2), is that it provides a coherent framework for studying uncertainties in both the data and model space. The term $\theta(d, m)$ is usually called *likelihood*, and ultimately represents the probability of observing the data, given particular model parameter values. As has been demonstrated in Section 4.3, this influence may be indirect in multiparameter and hierarchical models.

- $\mu_{D,M}(d, m)$ is a homogeneous distribution on $D \times M$

Posterior information on the model space is obtained by marginalizing (G.2) for m:

$$\Phi_m = \int_D \Phi(d, m)\, d\mathbf{d} = \rho_M(m) \int_D \frac{\rho_D(d)\theta(d|m)}{u_D(d)}\, d\mathbf{d} \tag{G.5}$$

The posterior distribution contains all interesting information on model parameters, including mean values, variance and confidence intervals for values of m.

Bibliography

[1] Metz J.A.J. et al. *The dynamics of physiologically structured populations.*, volume 68 of *Lecture Notes in Biomathematics*. Springer, 1986.

[2] Keyfitz N. *Applied mathematical demography.* Wiley, New York, 1st edition, 1977.

[3] Gurney W.S.C and Nisbet R.M. *Ecological Dynamics.* Oxford University Press, 1st edition, 1998.

[4] Kooiman S.A.L.M. *The dynamics of physiologically structured populations.*, volume 68 of *Lecture Notes in Biomathematics*, chapter Population dynamics on the basis of budgets. Springer, 1st edition, 1986.

[5] Perelson A.S. and Nelson P.W. Mathematical analysis of HIV-1 dynamics in vivo. *SIAM Review*, 41(1):3–44, 1999.

[6] Nowak M.A. and Robert M. *Virus Dynamics.* Oxford University Press, 1st edition, 2001.

[7] Ho D.D. Goudsmit J., De Ronde A. and Perelson A.S. Human immunodeficiency virus fitness in vivo: calculations based on a single zidovudine resistance mutation at codon 215 of reverse transcriptase. *Journal of Virology*, 70(8):5662–5664, 1996.

[8] Lloyd A.L. The dependence of viral parameter estimates on the assumed viral life cycle: limitations of studies of viral load data. *Proceedings of the Royal Society London*, 268:847–854, 2000.

[9] Goudriaan J. *The dynamics of physiologically structured populations.*, volume 68 of *Lecture Notes in Biomathematics*, chapter Boxcartrain methods for modeling of ageing, development, delays and dispersion. Springer, 1st edition, 1986.

[10] Jacquez J.A and Simon C.P. Qualitative theory of compartmental systems with lags. *Mathematical Biosciences*, 180:329–362, 2002.

[11] De Roos A.M. *Structured-population models in marine, terrestrial, and freshwater systems.*, chapter A gentle introduction to physiologically structured population models., pages 119–204. Chapman and Hall, New York, 1997.

[12] Sherman G. Jones S. Cohen S. Kuhn L. Hammer S. Loubser S., Balfe P. and Morris L. Sensitive real-time pcr quantification of 103N resistance mutants

following single-dose treatment with nevirapine. The 12'th conference on retroviruses and opportunistic infections., 2005. Personal communication with Loubser S., based on laboratory work performed at National Institute for Communicable Diseases (NICD).

[13] Tarantola A. *Inverse problem theory and methods for parameter estimation.* Elsevier, 1st edition, 1987.

[14] Carlin B.P. *Markov Chain Monte Carlo In Practice.*, chapter Hierarchical Longitudinal Modelling. Chapman and Hall, 1st edition, 1996.

[15] Gilks W.G. Spiegelhalter D.J., Best N.G. and Inskip H. *Markov Chain Monte Carlo In Practice.*, chapter Hepatitis B: a case study in MCMC methods. Chapman and Hall, 1st edition, 1996.

[16] Han C. and Chaloner K. Bayesian experimental design for nonlinear mixed-effects models with application to HIV dynamics. *Biometrics*, 60(1):50–59, 2004.

[17] Hu S. Banks H.T., Grove S. and Ma Y. A hierarchical bayesian approach for parameter estimation in HIV models. *Inverse Problems*, 21:1803–1822, 2005.

[18] Neumann A.U. Mittler J.E., Sulzer B. and Perelson A.S. Influence of delayed viral production in HIV-1 infected patients. *Mathematical Biosciences*, 152:143–163, 1998.

[19] Kirschner D. Using mathematics to understand HIV immune dynamics. *Notices of the AMS*, 43(2):191–202, 1996.

[20] Coombs D. Hyman J.M. Nelson P.W., Gilchrist M.A. and Perelson A.S. An age structured model of HIV infection that allows for variations in the production rate of viral particles and the death rate of productively infected cells. *Mathematical Biosciences and Engineering*, 1/2(2):267–288, 2004.

[21] Coombs D. Gilchrist M.A. and Perelson A.S. Optimizing within-host viral fitness: infected cell lifespan and virion production rate. *Journal of Theoretical Biology*, 229:281–288, 2004.

[22] De Lenheer P. and Smith H.L. Virus dynamics: A global analysis. *SIAM Journal on Applied Mathematics*, 63(4):1313–1327, 2003.

[23] Perelson A.S. Chen W. Leonard J.M. Ho D.D., Neumann A.U. and Markowitz M. Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature*, 373:123–126, 1995.

[24] Cao Y. Daar E.S. Ho D.D. Stafford M.A, Corey L. and Perelson A.S. Modeling plasma virus concentration during primary HIV infection. *Journal of Theoretical Biology*, 203(3):285–301, 2000.

[25] Cloyd M. Bajaria S.H., Webb G. and Kirschner D. Dynamics of naive and memory CD4+ t lymphocytes in HIV-1 disease progression. *Journal of the Acquired Immune Deficiency Syndrome*, 30(1):41–58, 2002.

[26] Paintstil E. Ricketts M. Gedzior G. Collins J., Thompson M.G. and Alexander L. Competitive fitness of nevirapine-resistant human immunodeficiency virus type 1 mutants. *Journl of Virology*, 78(2):603–611, 2004.

[27] Michor F. Iwasa J. and Nowak M.A. Virus evolution within patients increases pathogenity. *Journal of Theoretical Biology*, 232(1):17–26, 2005.

[28] Percus J. Coombs D., Gilchrist M.A. and Perelson A.S. Optimal viral production. *Bulletin of Mathematical Biology*, 65(6):1003–1023, 2003.

[29] Macdonald N. *Time Lags in Biological Models.*, volume 27 of *Lecture Notes in Biomathematics*. Springer, New York, 1st edition, 1978.

[30] Sompayrac L. *How pathogenetic viruses work.* Jones and Bartlett, 1st edition, 2002.

[31] Sompayrac L. *How the immune system works.* Blackwell Science, 1st edition, 2003.

[32] Metz J.A.J. and Diekmann O. *The dynamics of physiologically structured populations.*, volume 68 of *Lecture Notes in Biomathematics*, chapter Age Dependence. Springer, 1st edition, 1986.

[33] Wikipedia. RNA. http://en.wikipedia.org/wiki/RNA.

[34] Jones S. and Van Loon B. *Genetics for Beginners.* Icon Books, Cambridge, 1993.

[35] ABI Applied Biosystems. *ABI Prism SDS 7000 Software.* Applied Biosystems. ABI Prism 7000 Manual.

[36] Kainz P. The PCR plateau phase - towards an understanding of its limitations. *Biochimica et Biophysica Acta*, 1494:23–27, 2000.

[37] Kell D. Rowland J. Pritchard L., Corne D. and Winson M. A general model of error-prone PCR. *Journal of Theoretical Biology*, 234(4):497–509, 2005.

[38] Stern H.S. Gelman A., Carlin J.B and Rubin D.B. *Bayesian Data Analysis.* Chapman and Hall, New York, 1st edition, 1995.

[39] Gregory P. *Bayesian logical data analysis for the physical sciences.* Cambridge University Press, New York, 1st edition, 2005.

[40] Van Kampen N.G. *Stochastic Processes In Physics And Chemistry.* North-Holland Publishing Company, 1st edition, 1981.

[41] O'Hagan A. and Forster J. *Kendall's Advanced Theory Of Statistics.*, volume 2B of *Bayesian Inference.* Oxford University Press, 1st edition, 2004.

[42] Bernardo J. *Bayesian Theory.* Wiley, Chichester, 2000.

[43] Williams D. *Weighing the Odds; A Course in Probability and Statistics.* Cambridge University Press, 1st edition, 2001.

[44] Berger J.O. *Statistical decision theory and bayesian analysis.* Springer, 2nd edition, 1980.

[45] Richardson S. Gilks W.G and Spiegelhalter D.J. *Markov Chain Monte Carlo In Practice.* Chapman and Hall, 1st edition, 1996.

[46] Webb G.F. *Theory of nonlinear age-dependent population dynamics.* Marcel Dekker, New York, 1st edition, 1985.

[47] Caswell H. *Matrix Population Models.* Woods Hole Oceanographic Institution, 2nd edition, 2000.

[48] Karlin S. and Taylor H.E. *Stochastic processes: A second course.* Academic Press, 1st edition, 1981.

[49] Menke W. *Geophysical Data Analysis: Discrete Inverse Theory.*, volume 45 of *International Geophysical Series.* Academic Press, 1984.